Assessment of functional improvement in the hemiparetic arm following focal rehabilitation intervention

Ashford, Stephen

Awarding institution:
King's College London

Assessment of functional improvement in the hemiparetic arm following focal rehabilitation intervention

A thesis submitted to King's College London for the degree of Doctor of Philosophy

Stephen Ashford BSc MSc MCSP

Department of Palliative Care, Policy and Rehabilitation
School of Medicine
King's College London

January 2012

Abstract

The complex nature of upper limb function presents a challenge for rehabilitation following neurological injury. Some patients, with relatively mild injury, have potential to recover useful function such as the ability to use the hand to hold and manipulate objects (active function). Others with more severe injury will continue to have a non-functional upper limb, and may require assistance from another person (or their own non-affected arm) to care for the affected limb (passive function). The aim of this thesis was to develop and evaluate a self-report upper limb measure of active and passive function – the Arm Activity measure (ArmA) for use in focal spasticity management.

A systematic review demonstrated that no suitable measure was available, but provided possible items for inclusion in the ArmA. Patient-selected items were also included from goal setting for spasticity intervention. A modified Delphi consultation was undertaken to reduce the number of items, followed by item confirmation with a larger group of clinicians and pilot testing with patients and carers. The resulting twenty-item measure has two sub-scales of '*active'* and '*passive'* function.

Two inter-linked studies were undertaken, firstly to evaluate the psychometric properties of the ArmA, and secondly to undertake a hypothesis-generating cohort investigation of the course of functional changes following spasticity intervention.

Internal consistency evaluated by Cronbach's alpha was >0.85 for both sub-scales. Kappa coefficients for test-retest reliability were >0.90. Mokken analysis demonstrated unidimensionality for both subscales (Item H >0.5 for all items). Expected convergent and divergent relationships were seen with comparison measures (rho 0.5-0.63). The passive function sub-scale was responsive to change following spasticity intervention. In the cohort study, spasticity initially reduced following intervention and then increased again. Passive function improved and was maintained despite the increase in spasticity.

Adequate psychometric properties were demonstrated for the passive function sub-scale although further evaluation is indicated, particularly for the active function sub-scale.

Table of contents

Appendices

List of Tables

List of Figures

List of Boxes

Acknowledgements

This thesis arose from a clinical challenge with the management of spasticity in the hemiparetic upper limb for people with brain injury. Lynne Turner-Stokes was a pivotal influence in the development of the project, supporting me to develop it from initial concept to formal proposal.

My academic supervisors, Lynne Turner-Stokes and Mike Slade have provided immensely valuable and complementary support throughout the thesis. I would particularly like to thank them for their constant input despite the difficulties of the project and pressures of life in general. Lynne has provided suggestions at times for the direction of the project and Mike has challenged my conception of the work and methods required to do it. I would particularly like to thank them both for their focus not just on getting a project completed, but on developing me as an independent researcher.

I would also like to thank patients, carers and staff who helped in many ways with this work. In particular and in no specific order Ajoy Nair, Charlie Nyein, Fabienne Malaparade, Beverly Fielding, Hilary Rose, Sarah Stubbington, Ian Dolby, Denise Horn, Anita Jandrasec, Clare Belmont, Andrew Thu, Frances Clegg, Lisa Knight, Val Crook and Nila Shah. I would also like to thank my research colleagues and the team at the Department of Palliative care, Policy and Rehabilitation at King's College London. In particular Diana Jackson, Heather Williams, Jo Clark, and Richard Siegert. Heather Williams in particular has provided moral support and often acted as a 'sounding board' as well as providing assistance with double data entry and checking. Also, thanks to my mother Jane Ashford and brother Roland Ashford who have both provided comment on the final draft of the thesis and proof reading, for which I am extremely grateful.

Lastly, I would like to thank my wife Emma and children Samuel and Ella for their unfailing love and support over the course of this PhD. They have provided me with distraction when needed, wise council, reassurance, and I am indebted to them.

Publications related to this thesis

Peer-reviewed publications

Ashford S, Slade M, Malaparade F, Turner-Stokes L, (2008) *Evaluation of functional outcome measures for the hemiparetic upper limb – A systematic review*. **Journal of Rehabilitation Medicine**, 40 (10), 787-795. See Appendix 21.

Ashford S, Turner-Stokes L, (2009) *Management of shoulder and proximal upper limb spasticity using botulinum toxin and concurrent therapy interventions: A preliminary analysis of goals and outcomes*. **Disability & Rehabilitation**, 31 (3), 220-226. See Appendix 20.

Guidelines

Ashford S, (2009) Arm Activity measure (ArmA), Appendix 3, 54-55; In Royal College of Physicians, British Society of Rehabilitation Medicine, Chartered Society of Physiotherapy, Association of Chartered Physiotherapists Interested in Neurology. *Spasticity in adults: management using botulinum toxin - National Guidelines.* **Royal College of Physicians**, Editors: Turner-Stokes L, Ashford S,

Presentations at National and International Conferences

Ashford S, Slade M, Turner-Stokes L, (2008) *Preliminary evaluation of internal consistency and reliability of the arm activity measure (ArmA).* **Society for Rehabilitation Research**, Preston, UK, July 2 – 3, SRR proceedings (2008) Clinical Rehabilitation, 22, 856-863. Awarded the Verna Wright Prize for 2008.

Ashford S, Slade M, Turner-Stokes L, (2008) *Development of the Arm Activity measure for assessment of activity in the hemiparetic arm.* **5th World Congress for Neurorehabilitation**, Brasilia, September 24 – 27, Neurorehabilitation and Neural Repair (2008) 22, 514-640.

Ashford S, Slade M, Turner-Stokes L, (2008) *Evaluation of internal consistency and reliability of the arm activity measure (ArmA).* **5th World Congress for Neurorehabilitation**, Brasilia, September 24 – 27, Neurorehabilitation and Neural Repair (2008) 22, 514-640.

Ashford S, (2009) *Spasticity in Adults – Practice implications for using botulinum toxin. Including application and preliminary psychometric evaluation of the Arm Activity measure.* **Chartered Society of Physiotherapy Congress**, Liverpool, October 16-17.

Ashford S, (2009) *Goal Attainment Scaling - Benefits and pitfalls. Application of GAS in establishing the arm activity measure construct validity and responsiveness.* **Community Therapists Network Annual Conference**, Manchester, October 15.

Lannin N, McCluskey A, Cusick A, Ashford S, Ross L, (2010) *Measuring function in everyday life, enhancing the Disabilities of the Arm Shoulder Hand questionnaire for use post-stroke.* **World Federation of Occupational Therapy**, Santiago, May 4-7.

Ashford S, Slade M, Turner-Stokes L, (2010) *Development of the Arm Activity measure (ArmA) for assessment of activity in the hemiplegic arm.* **Chartered Society of Physiotherapy Congress**. Liverpool, October 15-16.

Ashford S, Nair A, Slade M, Turner-Stokes L, (2010) *Physical therapy and botulinum toxin-A (BTX), The relationship between spasticity reduction and passive function improvement.* **Chartered Society of Physiotherapy Congress**, Liverpool, October 15-16.

Ashford S, Turner-Stokes L, Slade M, (2010) *Psychometric evaluation of the Arm Activity measure (ArmA), a measure of active and passive function in the hemiparetic arm.* **Chartered Society of Physiotherapy Congress**, Liverpool, October 15-16.

Ashford S, Turner-Stokes L, Slade M, (2010) *Psychometric evaluation of the Arm Activity measure (ArmA) – a measure of active and passive function in the hemiparetic arm.* **American Academy of Physical Medicine and Rehabilitation**, Seattle, USA, November 5.

Ashford S, Turner-Stokes L, Slade M, (2010) *Physical therapy and botulinum toxin-A (BoNT-A) – The temporal relationship between spasticity reduction and functional gain***. American Academy of Physical Medicine and Rehabilitation**, Seattle USA, November 5.

Ashford S, Turner-Stokes L, Slade M, (2011) *The temporal relationship between spasticity reduction and functional gain following focal spasticity interventions***. World Confederation of Physical therapy**, Amsterdam, Netherlands, June 21.

Other presentations

Ashford S, *Seminar series on evaluation of outcome following botulinum toxin intervention for spasticity using Goal Attainment Scaling and the Arm Activity measure* (ArmA). October 6-10, 2008, Perth, Melbourne, Brisbane and Sydney, Australia.

List of Abbreviations

ABI	Acquired Brain Injury
ADL	Activities of Daily Living
ArmA	Arm Activity measure
BTX	Botulinum Toxin-A
CAHAI	Chedoke Arm and Hand Activity Inventory
CCR	Clinicians Categorisation of Response
CIMT	Constraint Induced Movement Therapy
COSMIN	COnsensus-based Standards for the selection of health status Measurement INstruments
CTT	Classical Test Theory
DASH	Disabilities of the Arm, Shoulder and Hand
ES	Effect Size
FDS	Flexor Digitorum Superficialis
FDP	Flexor Digitorum Profundus
FCU	Flexor Carpi Ulnaris
FCR	Flexor Carpi Radialis
FIM	Functional Independence Measure
FPL	Flexor Policis Longus
GAS	Goal Attainment Scaling
GRC	Global Rating of Change
ICC	Item Characteristic Curve
ICF	International Classification of Functioning Disability and Health
ICP	Integrated Care Pathway
IRT	Item Response Theory
LASIS	Leeds Arm Spasticity Impact Scale
LVM	Latent Variable Model
MAL	Motor Activity Log
MAS	Modified Ashworth Scale
MIC	Minimal Important Change
OT	Occupational Therapy
PCA	Principal Component Analysis
PT	Physical Therapy (not assigned to a specific discipline or profession)

PROMs	Patient Reported Outcome Measures
QUOROM	Quality of Reporting of Meta-analyses
ROC	Receiver Operating Characteristics
RRU	Regional Rehabilitation Unit
SAC	Scientific Advisory Committee
SEM	Standard Error of Measurement
SDC	Smallest Detectable Change
SRM	Standardised Response Mean
WHO	World Health Organisation

Glossary of terms

Active function	Where a functional task is performed by active movement of the individual's affected limb e.g. to reach for, grasp or manipulate objects.
Activity	The execution of a task or action by an individual.
Activity analysis	Identifies the demands of the activity including what activity, where it was performed, how often, how quickly and with what.
Activity limitations	Difficulties an individual may experience in involvement with performing activities.
Activities of daily living	Activities undertaken for personal physical care of an individual's body, such as showering or caring for ones physical and social needs.
Associated reactions	Involuntary mass patterns of muscle activation produced because of the upper motor neurone syndrome, and often related to effort.
Body functions	The physiological functions of body systems.
Body structures	The anatomical parts of the body such as organs, limbs, and their components.
Clonus	A rhythmic pattern of contraction occurring at a rate of several times per second, which can be demonstrated by a sudden stretch to the muscle.
Concatenation	Physical addition.
Co-contractions	Disrupted synergistic muscle activity often produced because of the upper motor neurone syndrome contributing to limb stiffness.
Deterministic	Matching exactly of the functional form of the item characteristic curve to the Guttman model.
Effect size	The ratio of the mean difference to the standard deviation of baseline scores.
Environmental factors	Environmental factors make up the physical, social and attitudinal environment in which people live.

Feasibility	The property of an outcome measure indicating it is suitable for routine clinical use.
Generalisability	The extent to which findings (from a study) can be generalised (or applied) to real-life.
Guttman scaling	Items are arranged in an order so that an individual who agrees with a particular item also agrees with items of lower rank-order.
Health outcome	The effect of intervention in relation to the achievement of an intended health goal.
Homogeneity	The unidimensionality of a set of items.
Idiographic	Term to describe measures that are developed from the perspective of the individual.
Impairments	Impairments are problems in body function or structure such as a significant deviation or loss.
Internal consistency	Refers to the interrelatedness of a set of items.
Interpretability	The degree to which qualitative meaning can be assigned to quantitative scores.
Interval	Interval data are data that have an equal distance between measurement points, but with no absolute zero. For example temperature in degrees Celsius.
Item Characteristic Curve	Describes the relationship between a latent trait and performance on an individual item.
Likert scale	A response scale usually with between five and seven points used in questionnaires e.g. strongly agree, agree, undecided, disagree, strongly disagree.
Measurement	Quantification of an observation against a standard.
Minimal important change	The smallest difference in score in the domain of interest which patients perceive as beneficial and would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management.
Muscle weakness	Reduced muscle power as a result of the upper motor neurone syndrome.

Nominal	Data that is simply placed into categories and can therefore be counted but not measured.
Nomothetic	Measures, which are useful in describing a population as a whole.
Ordinal	Ordinal data are data placed in rank order. For example, most preferred food to least preferred food.
Participation	Involvement in a life situation.
Participation restrictions	Problems an individual may experience in involvement in life situations.
Passive function	Where a task is carried out on the affected upper limb by the individual using the unaffected upper limb or by a carer e.g. cleaning the palm of the hand or armpit, cutting fingernails or positioning the arm.
Physical therapy	Physical interventions used in rehabilitation practice to affect the physical state.
Physiotherapy	The health profession of physiotherapy, using interventions often of a physical nature to positively affect people in need of rehabilitation or physical management.
Probabilistic	The probabilistic version of the Guttman curve allows small variations from the model, but requires that data still largely conform to the model.
Real-life	To evaluate function in the context of everyday activities.
Ratio	Ratio data are data that have an equal distance between measurement points and have an absolute zero. For example height or weight.
Responsiveness	The ability of an instrument to measure a meaningful or clinically important change.
Reliability	The degree to which a test measures the same attribute each time it is used.
Self-report	Self-administered and Patient Reported Outcome Measures (PROMs).

Sensitivity	The ability of an instrument to measure change in a state regardless of whether it is relevant and meaningful to the decision maker.
Sensitivity (predictive validity)	The proportion of actual positives correctly identified by a test when compared against actual outcome.
Spasms	Sudden involuntary, and often painful, movements. Precipitated by passive stretch, but may also be triggered by peripheral, noxious and visceral afferents.
Spasticity	Disordered sensory-motor control, resulting from an upper motor neurone lesion, presenting as intermittent or sustained involuntary activation of muscles.
Spastic dystonia	Abnormally increased muscle tone present at rest.
Specificity	The proportion of actual negatives correctly identified by a test when compared against actual outcome.
Standardised response mean	The ratio of the mean change (in a single group) to the standard deviation of the change scores.
Stochastic	Allowance for some variability in the shape of the Item Characteristic Curves functional form.
Unidimensionality	The degree to which all items in a measure evaluate aspects of the same construct.
Validity	The degree to which a test measures what it is intended to measure.
Ecological Validity	The degree to which the results in a study reflect the activities that actually occur in real-life. In addition, ecological validity is associated with generalisability.
Face validity	The degree to which the test looks as though it is measuring what the test is supposed to measure.
Concurrent validity	The comparison of the new measure with an existing gold standard measure both applied at the same time.
Construct validity	The degree to which a test measures the theoretical ideas underpinning a particular topic.

Content validity	The degree to which the test measures the overt manifestations of the theory.
Positive predictive value	The proportion of those identified by a test that actually have responded to intervention.
Negative predictive value	The proportion of those not identified by a test that have actually not responded to intervention.
Predictive validity	The degree to which responses to the test can predict future behaviour or events.

Chapter 1 Introduction

1.1 Clinical context

This thesis arose from a clinical challenge in the field of neurological rehabilitation. The challenge involved the management of spasticity in the hemiparetic upper limb and measurement of outcome in a manner meaningful to patients and their carers. Guidelines for the management of upper limb spasticity have highlighted the importance of measuring the impact of interventions at the level of improved function and care needs (Royal College of Physicians 2002; Royal College of Physicians et al. 2009). However, a need has been identified to re-conceptualise measurement of upper limb function from the perspective of patients and carers.

The manipulative function of the hand and arm is an ability, which sets humans apart from other primates. Humans use the upper limb for a wide range of complex activities, which include different combinations of grasping, twisting, pulling, pushing, reaching and lifting. The wide range of movement at the shoulder and elbow allow the positioning of the hand optimally for manipulation, and humans have developed a high level of fine motor control. Because of this complexity, when control of upper limb movement is lost, it is difficult to restore.

Neurological damage to the brain, for example as a result of stroke or trauma, typically leads to paralysis or weakness of the opposite side of the body (hemiparesis), which may be partial or complete. In the early stages after injury, the affected limbs are often flaccid (low-toned paresis), but after a few weeks muscle tone may start to return and can lead to the development of muscle over activity or 'spasticity'. Spasticity will often have unwanted effects, such as pain and result in secondary problems such as muscle stiffness and contracture. Even if return of active movement occurs, spasticity may still interfere with the fine motor coordination required for highly skilled tasks.

The complex nature of upper limb function presents a particular challenge for rehabilitation following neurological injury. Some patients, with relatively mild injury, have the potential to recover useful 'active function' of the hand and/or arm (see Box 1.1). Others with more severe injury will continue to have a non-functional upper limb,

and may require assistance from another person (or from their own non-affected arm) to care for the affected limb, for example to maintain hygiene, dress or support the arm. This has been called 'passive function' (see Box 1.1).

Box 1.1 Active and passive function

Active function: Where a functional task is performed by active movement of the individual's affected limb e.g. to reach for, grasp or manipulate objects.

Passive function: Where a task is carried out on the affected upper limb by the individual using the unaffected upper limb or by a carer e.g. cleaning the palm of the hand or armpit, cutting fingernails or positioning the arm.

(Sheean et al. 2010)

When evaluating the success of rehabilitation interventions such as spasticity management with botulinum toxin (BTX), the diversity of presentation poses a major challenge for outcome measurement. A further challenge arises from the need to evaluate function in the context of everyday real-life activities, as opposed to simply what is possible to observe during a clinic attendance. Self-reported outcome measures are gaining popularity as a means of capturing the impact of intervention for the individual in the context of their normal lives (Greenhalgh et al. 2005). They also offer the possibility to obtain follow-up information at a distance, and avoid unnecessary clinic visits.

This thesis centres on the development and initial evaluation of a self-report outcome measure for assessing both active and passive function in the hemiparetic upper limb – the Arm Activity measure (ArmA). The ArmA is then applied as an outcome measure in the context of a focal intervention for upper limb spasticity using botulinum toxin (BTX) and physical therapy (PT) interventions (see Glossary).

This introductory chapter will:

- Present a framework for describing the impact of health conditions at different levels of patient experience – the World Health Organisation (WHO) International Classification of Functioning Disability and Health (ICF).

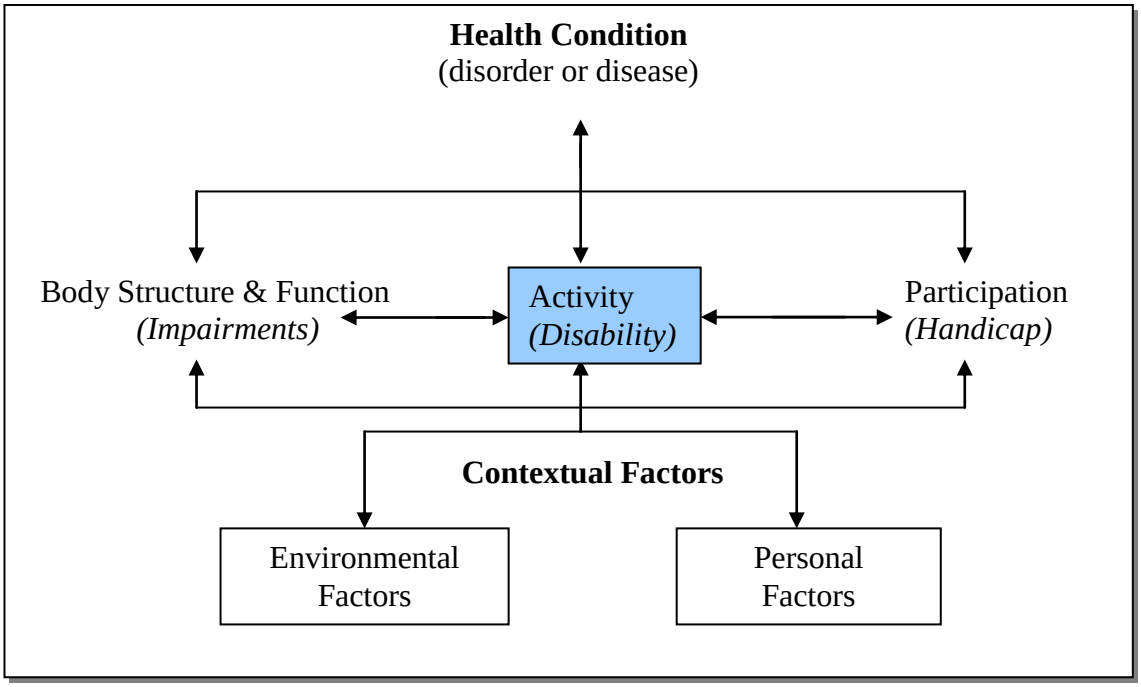
- Describe the effects of upper limb hemiparesis and the upper motor neurone syndrome in terms of that framework.
- Provide a brief overview of the clinical challenges in management and rehabilitation of the hemiparetic upper limb as it relates to the focal management of spasticity.
- Describe the physical interventions applied in rehabilitation of the upper limb, which have specifically been related to spasticity management.
- Explore the various measures currently used to assess the outcome of interventions within the activity level of the ICF.
- Examine the extent to which the existing literature has or has not demonstrated functional gains from such intervention, and explore some possible reasons for this.

1.2 Consequences of neurological damage

The International Classification of Functioning, Disability and Health (ICF) published by the WHO in 2001 can be used as a classification system to describe the consequences of neurological damage (WHO 2002). A further development of the original International Classification of Impairments, Disabilities and Handicap (ICIDH) (WHO 1980; WHO 2002) resulted in the ICF which classifies the consequences of disease as; impairment to body structure and physiological function, activity limitations (previously termed disability) and participation restrictions (previously termed handicap).

The ICF provides a model to describe the impacts of a health condition on a) the body, b) ability to perform activities, and c) participation in society within the context of personal factors relating to the individual and the environment in which they live and is shown in Figure 1.1 (WHO 2002; Stucki et al. 2007).

Figure 1.1 The ICF model of disability



(WHO 2002)

Application of the ICF provides a basis for rehabilitation practice and research with a standardised language to discuss the impact of disease incorporating participation, activity, and body systems. The World Health Assembly approved the ICF in 2001 and referred to rehabilitation in a resolution in 2005 (Stucki et al. 2007). The ICF has been widely used in rehabilitation practice and Stucki and colleagues have proposed that it provides a common framework to describe rehabilitation outcome, which is useful to both clinical practice and research (Stucki et al. 2007). It offers a common language to both communicate and describe the wider impact of disease processes. The category of activity is particularly important in rehabilitation where the primary outcomes of intervention are functional, and measurement at this level of the ICF is therefore needed. Table 1.1 shows examples of some of the common problems experienced by people with neurological damage affecting the upper limb, categorised using the ICF.

Table 1.1 Upper limb problems experienced by people with neurological damage

ICF domain	Problems experienced
Impairment to body structure and function	Paralysis Spasticity Contracture Pain
Limitation of activity	
Passive	Difficulty with caring for the limb: • Maintaining hygiene • Cutting fingernails • Dressing
Active	Unable to use the upper limb for active tasks: • Lifting, carrying • Reaching • Manipulating objects • Activities of daily living (ADL)
Restriction of Participation	Loss of employment Inability to engage in leisure activities

Whilst it is useful as a conceptual framework, the ICF contains some terminology that can be confusing. The ICF uses the term 'activity' to describe the impact of impairment on function for an individual. The term 'function' can be confusing in the context of the ICF because it is primarily used to refer to the 'function of the body' rather than activity. However, in common parlance function is principally used to describe activity. For the purposes of this thesis, the term function is used when referring to activity or aspects of activity unless otherwise specified. The ICF will be used as a conceptual framework to assist in understanding the context for measurement of upper limb function in hemiparetic individuals undergoing treatment for focal spasticity who have diverse presentations and goals.

1.2.1 Upper limb problems after stroke and brain injury

Upper limb impairment after central nervous system damage includes muscle weakness, spasticity, poor co-ordination and often sensory impairment (Geyh et al. 2004). Such impairments can result in functional restrictions to many activities of daily living (ADL; also see Glossary), such as washing, dressing or eating. Up to 70% of patients with stroke admitted to hospital, are identified to have arm weakness (Nakayama et al. 1994) and 60% have a "non functional" arm (Wade et al. 1983). Between 50% and 70% have long-standing restriction in arm function (Wade et al. 1983; Nakayama et al. 1994).

The upper motor neurone syndrome

The group of motor impairments resulting from damage to the central nervous system are termed the upper motor neurone syndrome (UMNS) (Thompson et al. 2005; Stevenson and Jarrett 2006). The UMNS is divided into positive and negative features (Pandyan et al. 2005; Thompson et al. 2005). Positive features are those characterised by muscle over activity and negative features by under activity (see Table 1.2).

Table 1.2 Positive and negative features of the UMNS

Positive features	Negative Features
Spasticity	Muscle weakness
Spastic dystonia	Loss of dexterity
Increased tendon reflexes	Fatigue-ability
Clonus	
Co-contraction	
Spasms	
Associated reactions	

(Pandyan et al. 2005; Thompson et al. 2005) (See Glossary)

The positive features of the UMNS are individually defined as follows: Spasticity as a clinical problem is often challenging to define and to manage (Pandyan et al. 2005; Thompson et al. 2005) but is characterized by involuntary muscle over-activity (see next section). Spastic dystonia is increased muscle activity at rest presenting as abnormal postures, such as the constantly clenched fist (Thompson et al. 2005). Co-contraction is needed for many normal movements, but becomes abnormal if both

agonist and antagonist contract with similar force and prevent movement (Stevenson and Jarrett 2006). Associated reaction, refers to involuntary activity of a body part, which occurs in response to voluntary movement elsewhere (Stevenson and Jarrett 2006). In patients with hemiparesis, this commonly occurs when the person is walking and the affected arm progressively flexes at the elbow. Clonus is a rhythmic pattern of contraction occurring at a rate of several times per second and can also occur with a sudden stretch to muscle (Stevenson and Jarrett 2006). Clonus is a positive feature of the UMN syndrome, which is not often a primary concern for management, particularly in the arm, but may be associated with other positive features.

Spasticity

Spasticity presents in a variety of ways depending on the size, location and age of the lesion, and may have associated unwanted effects such as pain, deformity and impaired function (Burke et al. 1988). Data on the prevalence of spasticity are varied, but it has been reported in 38% of patients at 12 months after stroke (Watkins et al. 2002), although Sommerfeld and colleagues reported prevalence as low as 19% at 3 months after stroke (Sommerfeld et al. 2004). Spasticity has been defined by Lance in 1980 as:

> "...a motor disorder, characterised by a velocity-dependent increase in tonic stretch reflexes (muscle tone) with exaggerated tendon jerks, resulting from hyper-excitability of the stretch reflex as one component of the upper motor neurone syndrome" (Lance 1980).

However, the SPASM consortium, a European thematic network to develop standardised measures of spasticity (Pandyan et al. 2005), have more recently proposed a broader definition:

> "Disordered sensory-motor control, resulting from an upper motor neurone lesion, presenting as intermittent or sustained involuntary activation of muscles"

This broader definition incorporates the positive features of the UMNS, but still excludes the negative features and biomechanical changes to muscle and associated structures. The authors argue that, in so doing, it is more clinically relevant.

Unfortunately, there remains no consistent and universally accepted definition of spasticity in use by researchers and clinicians. In practice, some clinicians and researchers refer to spasticity as hyperactive stretch reflexes, while others discuss a clinical presentation, which incorporates hyperactive stretch reflexes, other aspects of the UMNS as well as biomechanical changes to muscle and connective tissue. However, for the purposes of this thesis the broader SPASM definition of spasticity will be used.

If spasticity is not managed effectively, affected muscles adopt a shortened position with corresponding contribution to abnormal limb and trunk posture (Barnes 2003). This may result in soft tissue shortening and biomechanical changes in the contracted muscles (Pope 2002). Resistance to passive movement in muscle (following neurological insult) may be due to spasticity, thixotropy or contracture (Vattanaslip et al. 2000). Thixotropy is stiffness in muscle, which is dependent on the history of the limb movement (Vattanaslip et al. 2000). If muscle is not moved, there is a tendency for it to become stiffer and therefore more resistant to movement. If movement is not undertaken for long periods of time contracture will also develop. Contracture represents physical shortening of the muscle and other soft tissues around the joint (e.g. joint capsule), with loss of passive range of movement, which is irrespective of thixotropy or spasticity (Vattanaslip et al. 2000). However, contracture may also cause some stiffness towards the end of range of movement. Prevention of contracture and minimisation of thixotropy are therefore important in management of an immobile hemiparetic upper limb.

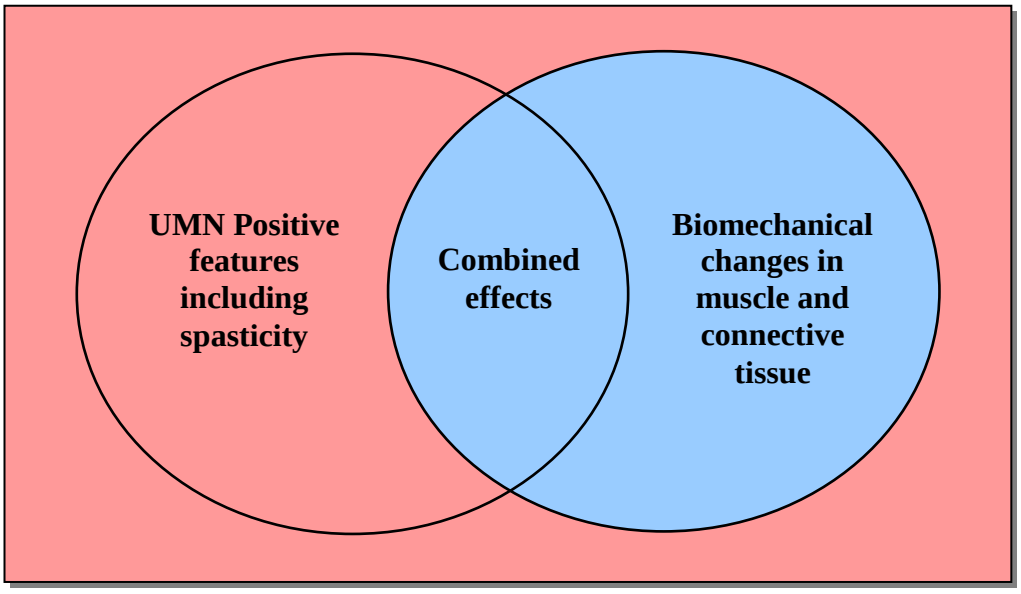
In addition, spasticity may also be painful for two reasons. First the over activity of the muscle itself can be painful (similar to ‘cramp’) because of the strength and duration of contraction (Stevenson and Jarrett 2006). Second, the resulting mal-alignment can cause pain by adversely stressing interconnecting muscle tissue and joint structures (Thompson et al. 2005). Figure 1.2 shows the clinical impact of arm spasticity at the shoulder, elbow, wrist and hand.

Figure 1.2 Clinical presentation of left arm spasticity.

The left arm is extended and adducted at the shoulder, flexed at the elbow, supinated at the forearm, with the fingers flexed into the palm of the hand.

Figure 1.3 represents the relative contributions of the positive features of the UMNS and biomechanical changes seen in the clinical presentation.

Figure 1.3 Resistance to passive stretch



1.2.2 Management of spasticity in the hemiparetic upper limb

Clinical management should address spasticity where it is causing harm to the patient's care provision; function, pain or risk of further deterioration exists, but should not treat spasticity indiscriminately without the possibility of meaningful benefit (Bhakta 2000b; Barnes 2003; Thompson et al. 2005). For example, the implication of spasticity in the hand may be in leading to contracture in the long finger flexors. The resulting difficulty in cleaning the palm of the hand may lead to maceration of skin tissue in the palm. Risk of secondary contracture and damage to the skin would warrant intervention to prevent this and management of spasticity is indicated. In contrast, if spasticity is present as a symptom without other adverse effects then intervention is not warranted.

At a clinical level, treatment will often be concerned with management of spasticity and other features of the UMN syndrome such as spastic-dystonia, co-contraction, and associated reactions (Stevenson and Jarrett 2006). The combined intervention package will also address other issues, such as joint range of movement (passive or active function) and task retraining (active function) (Thompson et al. 2005). In some patients who have active function goals, the negative UMN feature of weakness, will be the main issue addressed by both task training and specific strengthening (Thompson et al. 2005).

Spasticity may be focal (localised to a specific anatomical area affecting one or two muscle groups), regional (affecting the whole limb e.g. arm) or generalised (affecting the whole body) (Bergfeldt et al. 2006; Royal College of Physicians et al. 2009). The main approach to management has often been passive stretch, which is thought to both maintain structural length of muscle and inhibit expression of spasticity through an inhibitory effect on the stretch reflex (Thompson et al. 2005).

The evidence for physical interventions to manage contracture and muscle shortening is limited. The aim of such intervention is to counteract the dominant posture, which may lead to muscle and soft tissue shortening if unchecked. Passive movement, active movement or interventions such as splinting or casting can achieve muscle stretch. The possible advantage of splinting or casting is that they produce stretch of a longer duration than manual passive movement alone (Stevenson and Jarrett 2006; Lai et al. 2009).

However current reviews have not supported splinting and casting applications in the upper limb for maintaining muscle length particularly in the acute period immediately following insult (Katalinic et al. 2010; Tyson and Kent 2010). The primary sources of evidence in the reviews included all patients with a paretic limb, the majority of whom would not normally be considered for splinting intervention particularly acutely. The lack of difference between the intervention and control groups is therefore not as surprising as it first seems. In addition, an acknowledged weakness of a number of the included studies was that patients often did not receive the intended splinting dosage (Lannin and Herbert 2003; Lannin et al. 2007). The majority of trials reviewed in both of these systematic reviews were examining patients in the more acute phase after stroke and therefore differences may be seen in patients in a more chronic state. The effectiveness, methods and dosage required with physical interventions of this type are yet to be adequately identified and may also contribute to changing views on effectiveness (Katalinic et al. 2010).

Physical interventions may be sufficient to inhibit the development of contracture in muscle and soft tissue in some cases of upper limb spasticity (Stevenson and Jarrett 2006; Lai et al. 2009), although this is yet to be clearly demonstrated by research studies (Katalinic et al. 2010; Tyson and Kent 2010). However in moderate to severe spasticity, pharmacological treatment may be needed to support an effective management programme as an adjunct to physical intervention such as splinting (Stevenson and Jarrett 2006). For focal spasticity the pharmacological intervention of choice is intramuscular botulinum toxin (BTX) injection (Royal College of Physicians 2002; Royal College of Physicians et al. 2009).

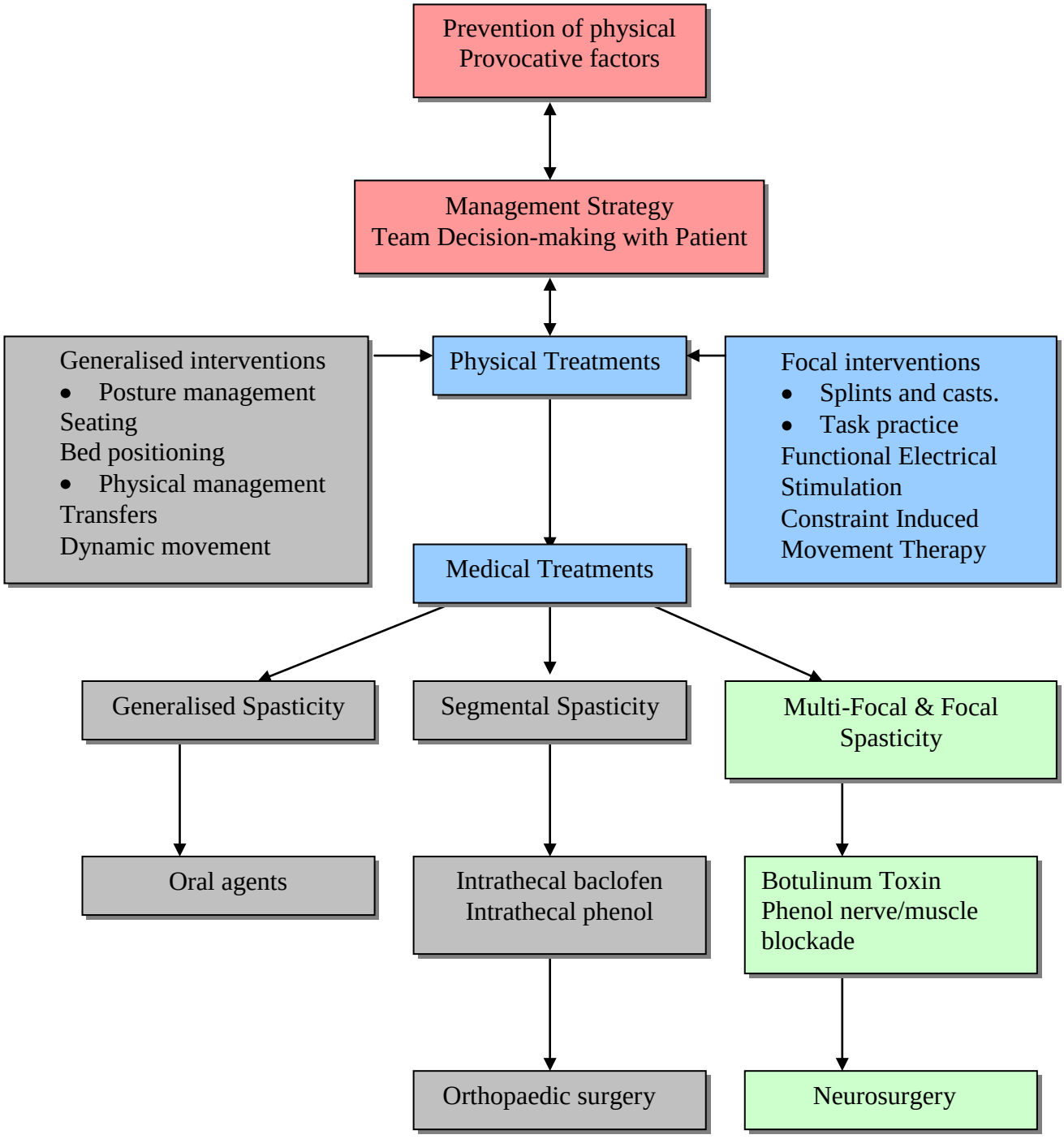
Generalised spasticity interventions will weaken all muscles, not just those that have spasticity. Therefore, focal interventions may also be desirable when:

a) Weakening of other muscles may lead to adverse consequences, such as trunk weakness effecting sitting balance, dysphagia or weakening respiratory muscles.
b) Muscle activity is wanted in opposing muscle groups to re-educate movement resulting in active function improvements.

Focal interventions are therefore particularly useful in these circumstances and the measurement of the functional effect of such intervention is a particular focus of this

thesis. Figure 1.4 summarises the overall management strategy and is adapted and updated from the Royal College of Physicians guideline document (Royal College of Physicians et al. 2009).

Figure 1.4 Strategy for management of spasticity in adults



Generalised and segmental interventions are included but 'greyed out' to emphasis the focal interventions, both physical and pharmacological, of interest in this thesis.

Botulinum toxin (BTX) intervention for focal spasticity

Focal interventions such as BTX have an advantage over systemic interventions in targeting specific muscles and are therefore less likely to increase weakness in other muscles (Stevenson and Jarrett 2006; Royal College of Physicians et al. 2009). In the minority of patients receiving spasticity intervention where improvement in active function is possible, reducing spasticity may restore a more normal balance of motor control and, in combination with task practice, may lead to improved active function.

BTX is a neurotoxin which acts by blocking transmission pre-synaptically at the neuromuscular junction (Barnes 2003; Dressler et al. 2005). The toxin is produced by *Clostridium botulinum* and strains of the bacterium have been identified as producing seven distinct toxins labelled A–G (Hambleton and Moore 1995). Toxin-A is the serotype that has been developed into a therapeutic agent and widely applied in clinical practice and research (Elia et al. 2009; Royal College of Physicians et al. 2009). Other serotypes have also been developed for use and botulinum toxin-B is available, but is less frequently applied in management of spasticity in the upper limb due to its shorter duration of action (Davis and Barnes 2000; Brashear et al. 2004).

BTX is used to manage spasticity by relaxing targeted muscles, inducing paralysis or partial paralysis. Reduction of spasticity allows stretching techniques such as splinting to reduce any biomechanical component affecting muscle and connective tissue length. The BTX itself is only effective in reducing the neurogenic component of spasticity (Royal College of Physicians 2002; Royal College of Physicians et al. 2009). There are therefore two key pre-requisites for the successful use of BTX in management of clinical spasticity:

- There must be a significant component of muscle spasticity, which can be reduced
- An appropriate programme of stretching including splint application, passive stretching program or movement re-education to maintain muscle length for passive function or retrain active function must follow injection (Royal College of Physicians et al. 2009).

Reduction in spasticity or other positive features of the UMN syndrome does not directly lead to functional improvement (Bakheit 2004a). It may allow passive function change to occur, by improving the ease of carrying out care tasks, because muscles and joints become easier to move. In some cases, active function may be improved by enabling muscle stretch and thereby facilitate re-education of movement (Royal College of Physicians 2002; Royal College of Physicians et al. 2009; Wissel et al. 2009). Alternatively, active function may potentially improve if spasticity is masking selective control. However reducing spasticity alone is unlikely to lead to improvement in function in the absence of other intervention (Sheean 2001; Royal College of Physicians et al. 2009).

Elia and colleagues conducted a recent systematic review and meta-analysis of BTX intervention for post stroke spasticity (Elia et al. 2009). They concluded that:

"Botulinum toxin A reduces upper limb spasticity in patients post stroke, but the improvement in functional ability remains to be established. This gap needs to be filled by new studies to assess the effect of BTX in the context of multidisciplinary patient management" (p 801).

There are a number of possible reasons for the failure to demonstrate functional improvement in clinical trials:

a) Lack of appropriate concomitant therapy to maximise the functional improvements (in many studies concomitant interventions are poorly described).
b) Failure to specify in advance the types of gains (passive or active) that are likely to be achievable – which will vary from patient to patient.
c) Lack of suitably focused measures to capture the gains that occur (passive and active).
d) Measuring at the wrong time points. Functional changes may take longer to develop than improvements in spasticity, because they depend on both reduction of spasticity and improvements resulting from physical interventions.
e) In the case of active function, failure to identify appropriate patients who may benefit.

Functional improvement following BTX is therefore not universally accepted and requires further exploration. In addition, the application of physical interventions (such as splinting) in combination with BTX, the specific roles these interventions have and the resulting improvements achieved also need more investigation.

Physical therapy (PT) intervention in combination with BTX

PT is used in this thesis to describe physical intervention used in rehabilitation practice to affect the physical state and is not used to refer to a profession. A small number of studies have explored the use of PT interventions in combination with BTX. The following section will describe the key findings of these studies.

In many studies of BTX intervention (discussed subsequently in Section 1.3), physical interventions are not included or described and the focus for evaluation is on reduction in spasticity rather than functional outcome. Some studies have attempted to evaluate the influence of physical interventions in this context. Giovannelli and colleagues demonstrated in a single-blind randomised controlled trial of patients with multiple sclerosis, that patients who received physiotherapy in combination with BTX had a significantly greater reduction in spasticity measured on the Modified Ashworth Scale (MAS) than those receiving BTX alone (Giovannelli et al. 2007). Unfortunately, the study did not evaluate functional outcome. In addition, the specific PT interventions used in this study were not described and other limitations included a small sample size, incomplete blinding and measurement bias. Despite the limitations of this study, it provides indicative support for a combination of BTX and PT intervention rather than BTX alone.

Other studies have attempted to explore specific PT interventions used in combination with BTX. Hesse and colleagues conducted a small randomised, double-blind, placebo-controlled trial investigating BTX in combination with functional electrical stimulation (FES) (Hesse et al. 1998). Four groups were compared; BTX, FES, BTX + FES and no intervention. Spasticity was measured using the MAS. They also measured limb position and difficulties encountered by the patient or carer in performing care tasks (passive function) such as cleaning the palm, cutting finger nails and putting the affected arm through a sleeve. Improvements were most prominent for spasticity reduction in the group who received combined BTX and FES. Of the functional tasks,

only ‘cleaning the palm of the hand’ showed significant improvement, which occurred in both the BTX + FES group and the FES only group. These results suggest that specific PT intervention such as FES may have a role in combination with BTX, but also that FES in some cases may be an effective management alone, to improve passive function at least in some patients.

Carda and Molteni (Carda and Molteni 2005) compared BTX and strapping with BTX, splinting and electrical stimulation. Both groups showed significant reductions in spasticity (as measured by the MAS), but a significantly greater reduction was identified in the BTX and strapping group. Again, this study used MAS as the primary outcome measure and did not attempt to measure functional outcome in any way. This study had significant limitations in the way in which intervention (strapping, splinting and electrical stimulation) was applied (with between group differences), control of bias (provision of intervention and measurement by the same clinician) and lack of blinding. Findings are therefore of more interest because of the questions they raise rather than the conclusions that can be taken forward and applied to practice.

All these studies had limitations in evaluation of these physical interventions including study design, limited sample size and lack of blinding. Overall, findings provide very limited evidence that focal PT interventions (such as splinting, strapping and FES) should be considered in conjunction with BTX in the management of spasticity and associated positive features of the UMN syndrome. Interventions used often need to be tailored to the specific needs of individual patients, which makes evaluation more challenging (Ashford and Turner-Stokes 2006; Ashford and Turner-Stokes 2008; Turner-Stokes et al. 2010). The trials of physical interventions used in combination with BTX have often focused on treatments that are easily available, rather than interventions, which may be more optimal. Further work is required to identify the most effective interventions, as well as dose requirements and methods of application, before recommending clear strategies for clinical practice.

1.3 Measurement of upper limb function following focal spasticity intervention with botulinum toxin

The ICF can be used as a system to classify measures of rehabilitation outcome. As discussed in Section 1.2, the four broad classifications of the ICF describe the impact of a health condition on the person. Each of these classification categories contains sub-classifications called domains. An example of a domain would be self-care; the ability to care for oneself. Within each domain, further classifications are provided, which have numeric codes. These classifications are termed category codes (Australian Institute of Health and Welfare 2003). Outcome measures can be mapped onto and then evaluated against these classifications, to identify the areas of the ICF, which they measure (Poissanta and Mayob 2004). Environmental factors can also be recorded as being either barriers to, or facilitators of, a person's activity (Australian Institute of Health and Welfare 2003). The ICF and more commonly its precursor the ICIDH have been used as reference frameworks in the development of a number of upper limb measures of activity. For example the Frenchay Arm Test (DeSouza et al. 1980), which measures upper limb activity in the clinic setting.

Global measures of function are widely used to measure improvement in self-care. Examples include the Modified Barthel Index (Wade and Collin 1988) and Functional Independence Measure (Keith et al. 1987). However these measures are often unresponsive to changes following focal intervention particularly in the upper limb, because localised changes, may be lost amongst the larger number of unchanging items (Granger et al. 1993). Conversely many of the specific measures e.g. Nine Hole Peg Test (Wade 1992b) or Action Research Arm Test (Carroll 1965) used to evaluate improvements in the upper limb, do not represent what patients do during a 'normal' week because they evaluate function in the context of a clinical examination rather than the normal environment (Jones 1990). Playford emphasises that many standardised measures, even if designed for specific anatomical areas such as the upper limb, do not capture how patients actually perform functionally and are often not reflective of the aims of intervention (Playford 2008).

The framework given by the ICF provides a basis for the exploration of function at the level of activity. Section 1.3.1 will explore the meaning of activity for patients

undergoing rehabilitation and will then address the need and possible methods of assessment of activity in clinical and research environments.

1.3.1 Assessment of activity

Assessing activity has two broad challenges, which are:

a) What exactly should be measured (i.e. what constitutes the activity of interest)? and
b) How should the activity be measured?

Activity is, by definition, very broad and therefore difficult to define. However, in the upper limb, related to focal spasticity intervention, this becomes easer to identify in terms of active and passive function.

Measuring activity in a clinical context also requires a balance between the accuracy of information obtained and the practicality of collecting the data. Three broad methods exist for the collection of activity data:

- Firstly, the observation of activity in the everyday environment,
- Secondly, observations of a proxy task such as that performed in a clinic situation and
- Thirdly, self-report of activity in the everyday environment.

Each method has different 'trade-offs' between the representative nature of the data produced and practical restrictions in measurement. The following section will briefly explore activity analysis, measurement in the clinic environment and then focus on self-report measures.

Direct observation of activity over a 24 hour period would be the most accurate way of determining what an individual actually does, but it is very time consuming and impractical for most clinical situations. Deconstruction of an activity (termed "activity analysis") identifies the demands of the activity including what activity, where was it performed, how often, how quickly and with what (Blake and Fritz 1996). Activity analysis requires either direct observation of the activity or methods such as video recording the tasks performed. It provides an accurate measure of function in the day-to-day environment (Blake and Fritz 1996).

Activity analysis has three principal limitations.

- Firstly, it is time consuming because an observer must be present with the individual or to review the video recording.
- Secondly, it is expensive to undertake, because of the cost of the observer's time.
- Thirdly, the presence of the observer may influence the function being performed.

These limitations result in activity analysis, while being a useful research tool, being impractical to apply in clinical practice.

Observation of a proxy task in the clinic environment is used in many existing upper limb outcome measures, for example the Frenchay Arm Index (DeSouza et al. 1980). Such measures require patients to complete standardised tasks, thought to be representative of the activities performed. Proxy measures applied in a clinic setting circumvent some of the problems of activity analysis (Berglund and Fugl-Meyer 1986; Wade 1992b; Blake and Fritz 1996; Alon et al. 1998; Lagalla et al. 2000), but may not be representative of everyday performance. They are usually more practical to apply, but will still have time costs to clinicians, carers and patients.

One approach to address some of these limitations is the use of self-report measures, which are also referred to as 'self-administered' and 'Patient Reported Outcome Measures' (PROMs). 'Self-report' means that the patient (and in some cases the carer) report on the activities, or aspects of them, that they have carried out in the home environment. This approach is more representative of real life, but carries the risk of inaccurate reporting. Self-report measures have the advantage of reporting on events that the clinician is not able to observe and are possibly less time consuming and expensive. In addition, reporting over longer periods than a single assessment point in the clinic reflects activity performed over this whole period, rather than at only one clinic appointment, which enhances ecological validity.

The use of self-report measures is commonly assumed less time consuming for clinicians. However, patients with cognitive and communicative impairments may have considerable difficulty independently completing the measure. Strategies are available to evaluate this ability in these patient groups and identify when simple questionnaire

completion may still be possible with support (Turner-Stokes and Jackson 2006). Other methods are also available to support those who may be unable to complete questionnaires due to communication impairment or milder cognitive limitations (Jackson et al. 2006). In practice, a combination of completion by patient and carer may be the most accurate form of feedback for care tasks particularly when both parties are involved in the process. Self-report measures do not avoid the problems inherent in all questionnaires, such as the time costs in data entry and dealing with missing data.

In summary, while activity analysis provides a gold standard measure of activity function, it is impractical for routine clinical practice due to time and cost constraints, which apply even in research. This thesis attempts to balance the need to measure activity in an ecologically valid manner, while ensuring feasibility (see Glossary) and a reduced burden for routine clinical practice. The following section explores measures applied in the evaluation of BTX intervention. Measures, which are patient focused or self-report, are then explored in more detail.

1.3.2 Patient and carer report in upper limb spasticity intervention

The case for appropriate measures to evaluate passive function has been made by Sheean (Sheean 2001) and is supported by the work of Bhakta (Bhakta et al. 1996; Bhakta et al. 2000a) and Brashear (Brashear et al. 2002; Brashear et al. 2004). These authors have attempted to identify patient reported methods to measure passive function outcomes following BTX intervention. Passive function has been evaluated in a small number of studies (Bhakta et al. 2000a; Brashear et al. 2002). However there is an identified gap in the measures available for the assessment of passive function in the hemiparetic upper limb, combined with a need to measure everyday performance in passive and active function.

Many of the trials of BTX intervention have used the MAS as the primary or sole outcome measure (see Table 1.3). To characterise existing approaches to evaluation, the literature is presented in Table 1.3, showing measures used to evaluate outcome in trials of BTX intervention for management of spasticity. When PT interventions are mentioned, these are also presented, however descriptions are often limited. In addition, intervention and outcome are also described.

Table 1.3 Outcome measures in trials of BTX for upper limb spasticity

Author Year	Design	Subjects	Intervention	Outcome Measures		Findings
				Impairment	Activity	
(Simpson et al. 1996)	Multicentre PC- RCT 3-mth follow-up	N=39 Mixed CNS disease	Botox 75,150,300u Elbow and wrist flexors	Modified Ashworth Scale (MAS) Global Assessment of Response	Not measured	Significant improvement in spasticity and physician and patient Global Assessment of Response at 4 and 6 weeks.
(Yablon et al. 1996)	Open label study 2-4 wk follow-up	N=21 TBI	Botox Flexible regime (100-300u) Wrist/fingers followed by passive range of motion (ROM) exercise, splinting and casting as clinically indicated.	MAS Range of Movement (ROM)	Not measured	Significant improvement in spasticity and ROM in the distal upper extremity.
(Bhakta et al. 1996)	Open label study	N=17 Stroke	Botox or Dysport Elbow/wrist Variable dosage	MAS ROM (Goniometer) Pain (initial VAS)	Upper limb Function items identified by patients (e.g. hygiene, dressing, standing & walking balance)	Significant improvement in MAS and ROM. Reported functional improvement in 14/17. Safe and effective treatment for reducing spasticity.
(Hesse et al. 1998)	PC-RCT of BTX plus FES	N=24 Four groups 6 in each Stroke	Dysport 1000u into 6 upper limb muscles	MAS Limb position	3 functional tasks (palm hygiene, cutting finger nails, and arm through sleeve)	BTX and FES showed significantly greater reduction in difficulties with palm hygiene, with trend towards lower spasticity. The study had insufficient power.
(Bakheit et al. 2000)	Multicentre PC-RCT Dose-ranging study 4-mth follow-up	N=82 Stroke	Dysport 500,1000 or 1500 5 muscles– fixed regimen	MAS ROM (Goniometer and Scale for hand) Pain (scale)	Barthel Index Patient report items (palm hygiene, cutting finger nails, and arm through sleeve) Rivermead motor assessment (arm scale-RMA)	BTX A reduces spasticity (MAS) over 16 weeks (all doses). Optimal dose 1000U Dysport Intervention safe. No difference on RMA or Barthel Index between groups. Trend towards effect on functional tasks.

Author Year	Design	Subjects	Intervention	Outcome Measures		Findings
				Impairment	Activity	
(Bhakta et al. 2000a)	PC-RCT 3-mth follow-up	N=40 Stroke	Dysport 1000u Flexible upper limb regimen	MAS, Strength (Medical Research Council, MRC) Grip strength (Dynamometer) ROM (Goniometer) Pain (Numeric scale)	8-item Patient Disability Scale* 4-item Carer Burden Scale*	Significant improvement in patient disability at weeks 2-6, but reducing by week 12. Improvement in carer burden maintained at 12 weeks. Significant reduction in spasticity at 2 weeks but not maintained to 6 and 12 weeks. Some improvements in ROM. No group differences in strength or pain.
(Richardson et al. 2000)	PC-RCT 3-mth follow-up	N=52 32upper limb 20lower limb Mixed CNS disease	Botox Flexible regimen	Problem severity MAS ROM (Goniometer)	Goal attainment Rivermead Motor Assessment (RMA - arm, trunk and leg scales) 9-hole peg test, 10m walk	Significant improvements in MAS, ROM, problem severity and RMA in treatment group. Goal attainment achieved in both groups.
(Smith et al. 2000)	PC-RCT 3-mth follow-up	N=21 Stroke or TBI	Dysport Flexible dose regimen Elbow/wrist/fingers	MAS ROM (Goniometer) Posture	Global assessment scale Upper body dressing time Frenchay Arm Test	Reduction in spasticity and increased ROM in wrist and fingers at 6 weeks – lost by 12 weeks. No change in disability, but improvement on patient global assessment.
(Bakheit et al. 2001)	Multicentre PC-RCT 4-mth follow-up	N=59 Stroke	Dysport 1000u Flexible regimen	MAS ROM (Four point scale) Pain (Four point scale)	Patient report items (palmar hygiene, cutting finger nails, and arm through sleeve) Goal Attainment Scaling (GAS) Global benefit. Barthel Index (BI)	Significant reduction in spasticity up to 16 weeks. Significant improvement in elbow ROM at 16 weeks and global assessment of benefit. Tendency for patient reported items to show benefit – not formally analysed. No significant difference in pain, GAS or BI.
(Brashear et al. 2002)	Multicentre PC- RCT 3-mth follow-up	N=126 Stroke	Botox Flexible dose regimen Wrist/fingers	MAS	Disability Assessment Scale (DAS) Global Assessment Scale	Significant treatment effects on spasticity, functional ability (DAS) and global assessment.

Author Year	Design	Subjects	Intervention	Outcome Measures Impairment	Activity	Findings
(Childers et al. 2004)	Multicentre PC- RCT 6-mth follow-up	N=91 Stroke	Botox 90, 180, 360 U or placebo Fixed regimen	MAS	Functional Independence Measure (FIM) Global Assessment SF-36	Dose dependent reduction in spasticity but did not translate into improved function or quality of life.
(Carda and Molteni 2005)	Case-control study	N=65 Stroke	Botox Dose not specified Group 1: BTX + Strapping Group 2: BTX + Splinting and FES	MAS		Both groups improved significantly; but group 1 had significantly better reductions in spasticity than group 2 (MAS).
(Yelnik et al. 2007)	RCT	N=20 Stroke	Dysport 500u Subscapularis Non standard physical therapy	Pain (VAS) MAS (shoulder, elbow, wrist & fingers) ROM		Significant improvements in all measures in intervention group, apart from shoulder spasticity, which was unchanged.
(Giovannelli et al. 2007)	RCT Single blind	N=38 MS	BOTOX Upper limb 100U Lower Limb 100-300U Passive exercise and stretching (Experimental Group)	MAS		Significant reduction in spasticity, both groups. Greater effect in the experimental group.
(Kong et al. 2007)	RCT Double blind	N=17 Stroke	Dysport 250u Pectoralis Major; 250u Biceps Brachii Exercises following injection	Pain (VAS) MAS Shoulder range of movement		Significant reduction in shoulder spasticity (MAS). No significant change in shoulder pain (VAS).
(Marco et al. 2007)	RCT Double blind	N=29 Stroke	Dysport 500u Pectoralis Major for shoulder pain Fixed regimen TENS for 6 weeks	Pain (VAS) MAS Shoulder range of movement		Significant reduction in shoulder pain following BTX administration. Improved external rotation.

Author Year	Design	Subjects	Intervention	Outcome Measures		Findings
				Impairment	Activity	
(Jahangir et al. 2007)	RCT Double blind	N=52 Stroke	Botox 80u (20u each into FCU, FCR, FDS, FDP) 1 hour physiotherapy x 2 weekly	MAS	Barthel Index Euroqol EQ-5D	Significant reduction in MAS at wrist and fingers. No difference detected with Barthel Index and Euroqol.
(Elovic et al. 2008)	Cohort study	N=279 Stroke	Up to 5 treatments of (200-400 units) of botulinum toxin-A (BOTOX® wrist, finger, thumb, and elbow flexors) Divided into high and low dose groups.	MAS	1) Disability Assessment Scale (DAS) 2) Stroke Adapted Sickness Impact Profile (SA-SIP30). 3) Visual Analogue Scale from European Quality of Life-5 (EQ-5D)	Significant improvement in DAS – patients principal therapeutic target and spasticity at all time points, but not for low dose at the elbow (<250 units) and not for thumb at 12 and 24 weeks (<250 units). Significant changes in SA-SIP30 from baseline (total score).
(Bhakta et al. 2008)	PC-RCT 3-mth follow-up	N=40 Stroke	Dysport 1000u Flexible regimen	Biomechanical and electromyogram (EMG) measure of upper limb associated reaction.	Goal attainment scale using 10-point categorical scale for daily activities.	Significant improvement in biomechanical measure of associated reaction. No significant reduction of the interference of associated reactions with daily activities reported by patients.
(Chang et al. 2009)	Prospective cohort study	N=14 Stroke	BTX-A Repetitive task practice (60 mins per day) FES (Ness H200 system)	MAS	Chedoke-McMaster Assessment (CMA) at baseline for group allocation to; high 'function' group; lower 'function' group Action Research Arm Test (ARAT) Motor Activity Log (MAL)-28 MAL-5	Primary and secondary outcomes improved significantly over the 12-week intervention period. No differences between groups for ARAT and MAS. High function group predicted by CMA had significantly better MAL-28 outcomes than low function group.
(Kanovsky et al. 2009)	RCT Parallel group Follow-up to 20 weeks.	N=148 Stroke	Group 1; Botulinum toxin (Xeomin®) (median 320 units) wrist and finger flexors. Group 2; placebo group	MAS	Disability Assessment Scale (DAS) Global assessment of outcome	Significantly higher proportion of experimental group responders (> or = 1 point on MAS) at 4 weeks. Significant results for experimental group until week 12 in patient's principal therapeutic target (DAS), global disability and some care tasks.

Author Year	Design	Subjects	Intervention	Outcome Measures Impairment	Activity	Findings
(Lai et al. 2009)	RCT parallel group Follow-up at 1 and 14 weeks	N=30 Stroke	All patients received BTX-A, heat, education, mobilisation, passive and active range of movement stretch, proprioceptive neuromuscular facilitation and therapeutic exercise. Experimental group: elbow extension using Dynasplint®	1. Elbow active range of motion. 2. MAS		Percentage difference improvements in both measures in both groups were seen. Greater changes were seen in experimental group however; significant difference between the groups was not reported.
(Meythaler et al. 2009)	RCT crossover Follow-up to 24 weeks	N=21 Stroke	BOTOX® 300-400U with a therapy programme or placebo with a therapy programme.	MAS Deep tendon reflexes Passive and active range of movement MRC scale Grip (Dynamometer)	MAL Klein-Bell Activities of Daily Life scale MOS-36 Item Short-Form Health Status Survey	Significant difference from control in MAL (quality of movement sub-scale). No difference on other measures. Both groups showed reduction in MAS at 6 weeks with a return to baseline at 12 weeks.
(McCrory et al. 2009)	PC-RCT parallel group	N=96 Stroke	Dysport (750-1000u) or placebo Distal upper limb muscles 2 occasions, 12 weeks apart Flexible regimen Routine Physiotherapy	MAS Hospital Anxiety and Depression Rating Scale (HADS) Pain (VAS)	Assessment of Quality of Life (AqoL) Modified Motor Assessment Scale (MMAS) GAS 8-item Patient Disability Scale* 4-item Carer Burden Scale* Rating global benefit	Intervention and control groups did not differ according to primary outcome (AQoL). Significant improvements for GAS, MAS and improved global benefit. Modest reduction in patient disability and carer burden.
(Simpson et al. 2009)	RCT parallel group Review at 3, 6, 12 and 18 weeks.	N=60 Stroke	1. Botulinum toxin-A (BOTOX®) plus oral placebo 2. Tizanidine plus intramuscular placebo 3. Intramuscular and oral placebo	MAS Wrist flexor, elbow and fingers	DAS Modified Frenchay Scale	BOTOX® produced a significantly greater reduction in MAS than Tizanidine or placebo at 3 and 6 weeks. DAS showed a significant improvement in cosmesis at 6 weeks.

Author Year	Design	Subjects	Intervention	Outcome Measures		Findings
				Impairment	Activity	
(Barnes et al. 2010)	parallel group RCT	N=192 Stroke	Botulinum toxin (Xeomin®) 2 Groups 1. 50 Units 2. 20 Units	MAS	DAS	At four weeks (post injection) 1 point or greater reductions in DAS and MAS for both groups. Difference between responders and non-responders in groups for DAS. 10.6% greater response in group 1 (50 units).
(Cousins et al. 2010)	parallel group RCT	N=30 Stroke	Botulinum toxin (BOTOX®) – dosing based on Royal College of Physicians (2009) guideline Three groups 1. Half dose 2. Quarter dose 3. Placebo	1) Spasticity (sEMG) 2) Grip strength (JAMAR Dynamometer) 3) Isometric muscle strength – elbow and wrist	Action Research Arm Test (ARAT)	No benefit treatment over control in whole group analysis. Sub-group analysis (no active function ARAT, baseline), both active groups improved compared to control (effect size 0.5 – quarter dose and 0.6 half dose groups).
(Kaji et al. 2010)	Parallel group RCT	N=109 Stroke	Botulinum toxin (BOTOX®) Group 1 high dose (200-240 units); Group 2 high dose placebo group; Group 3 low dose (120-150 units); Group 4 low dose placebo group	MAS	Disability Assessment Scale (DAS)	Significant difference for mean difference (area under curve) groups 1 to 2 (high dose). No difference groups 3 to 4 (low dose). Significant difference noted in DAS at weeks 6, 8 and 12 groups 1 and 2 (High dose).

Author Year	Design	Subjects	Intervention	Outcome Measures		Findings
				Impairment	Activity	
(Shaw et al. 2010)	Open-label, parallel-group RCT	N=333 Stroke	Dysport (up to 1000u) 4 Week therapy programme (placebo group received therapy programme only) 2 Therapy programmes: 1) ARAT (0-3) - Stretching 2) ARAT (4-56) - Stretching and task training	MAS	ARAT 9-hole peg test Barthel Index Motricity Index (MI) Patient report items (palmar hygiene, cutting finger nails, and arm through sleeve) (Participation measures also recorded: Stroke Impact Scale European Quality of Life-5 Dimensions Oxford Handicap Scale)	No significant difference on primary outcome (ARAT). MAS reduced at 1 month in experimental group, but not 3 or 12 months. Change in strength (MI) at 3 months. Significant improvement in patient report items.
(Sun et al. 2010)	Parallel group RCT Review at 1, 6 and 6 months.	N=32 Stroke	Dysport® 1000u elbow and wrist flexors in both groups. Experimental group received CIMT with graded tasks, verbal feedback and a behavioural contract for 5 hours/ day for 3 months. Control group received conventional (neurodevelopmental) therapy for 2 hours/3 days per week for 3 months.	MAS	MAL ARAT	Both groups demonstrated significant improvement in spasticity (MAS) 1 and 3 months following injection. Significant difference between experimental and control group for elbow, wrist, and finger spasticity at 6 months post injection. Significant difference in MAL (amount of use score) and ARAT at 3 and 6 months.

Key: *Now known as the Leeds Arm Spasticity Impact scale (LASIS) – see Systematic Review; Chapter 5 for more information; PC = placebo controlled; RCT = randomised controlled trial.

Many of the trials addressing focal spasticity intervention with BTX have demonstrated significant improvements in the MAS (see Table 1.3). However the use of MAS alone is inappropriate because it does not reflect the primary aim of intervention (Bakheit 2004a). The reason for implementing the intervention is usually functional and often passive function (Bakheit 2004a). An option suggested by Bakheit is using the ICF as a model to interpret the goals of intervention and matching the outcome measures to goals at the activity level of the ICF. Using the ICF as a model to interpret goals has already been applied in spasticity management (Turner-Stokes et al. 2010).

A relatively small number of the trials presented in Table 1.3 attempt to evaluate active and passive function. Where studies have applied measures of global disability or function these have not shown significant change e.g. (McCrory et al. 2009). However when specific measures of disability/function or goal attainment have been applied, significant improvements in passive function have been demonstrated in some cases (for example studies by Bhakta and colleagues (Bhakta et al. 2000a) and Turner-Stokes and colleagues (Turner-Stokes et al. 2010)).

The remainder of this section describes the different patient focused methods used to evaluate functional outcome following focal spasticity intervention using BTX. The methods discussed are: the use of patient report on upper limb items (including the Leeds Adult Spasticity Impact Scale - LASIS), the Disability Assessment Scale (DAS), use of a composite measure of function and Goal Attainment Scaling (GAS).

Patient report - upper limb function items

Specific items thought to represent upper limb passive function have been used in some studies to measure functional outcome and are discussed in the following section. Three upper limb function items have been used in trials evaluating BTX intervention in a study by Hesse and colleagues (Hesse et al. 1998) and another by Bakheit and colleagues (Bakheit et al. 2000). The three items used were: 1. putting the arm through a sleeve, 2. cleaning the palm and 3. cutting fingernails. Each item is scored on a five-point Likert scale from “no difficulty” to “not able to do the activity”. The items are presented as a structured interview in the spasticity clinic setting. In the study by Bakheit and colleagues, improvement was demonstrated in the MAS, but was not shown in these functional items (Bakheit et al. 2000). However in a small study, Hesse

showed a significant reduction in difficulty with palmar hygiene (Hesse et al. 1998). Use of these patient reported items demonstrates some potential for them to register meaningful change following focal spasticity intervention.

Leeds Adult Spasticity Impact Scale (LASIS)

The LASIS has been developed as a more extensive measure of patient and carer reported functional outcome following BTX intervention for spasticity (Bhakta et al. 1996). Bhakta and colleagues undertook a preliminary evaluation of the impact of BTX treatment on disability caused by upper limb spasticity after stroke (Bhakta et al. 1996). In the literature, the measure is not referred to as the LASIS, but has been identified as such following contact with the author. A component of this study was the development of an 'item bank' following open interviews with patients to identify which care tasks were most important to them. The resulting measure has two sub-scales a Patient Disability Sub-scale and a Carer Burden Sub-scale. The Patient Disability Sub-scale assesses the patient difficulty in cleaning the palm, cutting fingernails, putting the paretic arm through sleeves, cleaning under the armpit, cleaning around the elbow, standing balance, walking balance and ability to perform a home exercise programme. The Carer Burden Sub-scale is used if a carer is involved, to assess cleaning the palm, cutting fingernails, dressing and cleaning the armpit. The items are scored either by the patient (Patient Disability Score) or carer (Carer Burden Score) on a scale between "no difficulty with task" to "unable to do task". The LASIS was delivered as a structured interview, based on the respondent's report on the preceding 7 days. Both sub-scales were summed and divided by the total number of completed items to give a separate Patient Disability Score and Carer Burden Score.

Bhakta and colleagues conducted a further study using this measure to investigate whether reduction in spasticity after BTX treatment translates into reduction in disability and carer burden (Bhakta et al. 2000a). Changes in Patient Disability Score occurred from 2 to 6 weeks but were not sustained to 12 weeks and changes in Carer Burden Score occurred by 6 weeks and were sustained at 12 weeks.

Disability Assessment Scale (DAS)

Brashear and colleagues (Brashear et al. 2002) developed the DAS, a measure of "functional impairment" associated with spasticity. Although it contains similar items

to those used by Bhakta and colleagues, it is scored by clinicians after a combination of interviewing and observing the patient. The measure assesses hand hygiene, dressing, limb position, and pain. All items are scored using a Likert scale. Each item is scored 0 - no disability, 1 – mild disability (noticeable but does not interfere significantly with normal activities), 2 – moderate disability (normal activities require increased effort and/or assistance), or 3 severe disability (normal activities limited). Brashear and colleagues (Brashear et al. 2002) undertook reliability testing and established the reliability of DAS and MAS in the same study. The DAS primarily addresses distal upper limb function, particularly of the wrist and hand. Although it includes an item on dressing, which will necessarily include shoulder movement, proximal movement of the arm is not fully addressed.

Brashear and colleagues (Brashear et al. 2002) then used DAS in a randomised controlled trial to evaluate if BTX reduced disability in persons with spasticity of the wrist and fingers after stroke. They concluded that intramuscular injection of BTX reduced spasticity of the wrist and finger muscles (according to MAS) and that disability in patients was also shown to decrease (according to DAS). Patients in the intervention group in this study showed a statistically significant improvement of at least one point on the DAS compared to the control group at six weeks after injection.

Strengths and limitations of LASIS and DAS

The scoring of LASIS and DAS are not truly ‘patient report’, because completion is undertaken by the clinician. Both measures are administered as structured interviews, with the addition of observation in the case of the DAS. However, the findings from evaluation of LASIS and DAS support the view that changes in disability and carer burden, following focal spasticity intervention, can be detected using patient or carer reported information. Although the changes in DAS were small, the results also indicate that BTX intervention for wrist and finger spasticity leads to functional improvement measured using patient reported information. Existing clinician-rated measures may detect change, but are insufficiently specific to fully characterise the improvement.

Both LASIS and DAS have limitations in terms of the incomplete evaluation of their scaling properties. The measurement properties of both of these measures have not

been confirmed. In order to undertake arithmetic manipulations to derive a total score, evaluation of measurement scaling properties is needed (also see Chapter 3; page 63). Before this, it is also necessary to confirm that a single dimension or construct is being measured or to identify the dimensions corresponding to sub-scales within the tools. Given that LASIS contains passive function items, but also some active function items it is possible that more than one sub-scale is needed or the rationalisation of items to form a single scale from a measurement perspective. Further work is therefore required in these measures before they are widely used in research and clinical practice.

In general, self-report arm function measures have addressed distal function as the primary concern. However, there is a need to ensure that measurement of upper limb function includes the contribution of the proximal upper limb in addition to distal and whole arm function.

Composite measure of function

The following section describes an approach taken to measuring function using a composite measure of function in a secondary analysis. A meta-analysis by Francis and colleagues (Francis et al. 2004) using data from two trials by Bakheit and colleagues (Bakheit et al. 2000; Bakheit et al. 2001) was undertaken using two composite scores. The aim of the study was to attempt to extract the data most relevant to the intervention and combine this, whilst excluding irrelevant items (e.g. non upper limb items of the Barthel Index). A composite spasticity score was generated, which comprised MAS scores from elbow, wrist and finger flexors and a composite functional score used three items from the Barthel index; feeding, dressing and grooming combined with the three other functional items (cleaning palm, cutting fingernails and putting the arm through a garment sleeve).

For the majority of patients in the meta-analysis change in function occurred at the same time as maximal change in spasticity at 4 weeks following intervention. Improvements were primarily seen in passive function. However, for a minority of patients, maximal functional change appeared to follow on at some time after maximal change in spasticity. The reason for the delay in these patients is unclear, but these findings may suggest that other factors such as time for PT intervention to have an effect is important in producing functional changes.

The authors suggest that future studies should have multiple time points and apply two primary outcome measures, a measure of spasticity (impairment) and a measure of function (activity). They also suggest using a measure of goal attainment such as GAS because of the diverse nature of the aims of treatment. Some limitations of the methods are identified, such as summing of scores in both the spasticity and functional indices, which may not be considered appropriate by some because of the non-parametric quality of the data. In addition, the functional index incorporates items from the Barthel index, which are active function (grooming, feeding and dressing), with the additional items likely to be passive function (cleaning the palm, cutting fingernails and putting the arm through a sleeve). It is therefore questionable whether these scores should be summed when they come from different constructs. Finally, the authors also concede that the time points used in the two studies included for meta-analysis were further apart than is ideal and that the exact time of change in spasticity and function may have been missed.

Goal Attainment Scaling

The Goal Attainment Scaling (GAS) approach attempts to evaluate the attainment of patient agreed goals of intervention rather than use a standardised set of measurement items. GAS is an individualised method, first introduced in the 1960s by Kiresuk and Sherman (Kiresuk and Sherman 1968). It is a system of evaluating achievement of the goals of intervention set by the patient and clinical team before the commencement of treatment. The method therefore provides both a quantification of individual goal outcome and qualitative information about the specific goals set. It is therefore an evaluation of expected goal achievement, dependent on the patient's ability to change and the clinical team's ability to predict that change. Because of its individualised approach, it does not provide a standardised outcome for comparison and therefore can not replace standardised measures.

Improvement in GAS is rated from –2 to +2. The expected target of achievement is set by the patient and treating team, before any intervention and given a value of 0. Outcomes less than expected are given values of –1 or –2 and more than expected +1 or +2. It is recommended by the originators of the method, that a 'T-score' is produced (Kiresuk and Sherman 1968). The 'T-score' is a total score of the composite outcome

of all the goals set for an individual patient (see Appendix 3). The validity, reliability and responsiveness of GAS have been evaluated in rehabilitation (Malec 1999) and upper limb focal spasticity intervention (Ashford and Turner-Stokes 2006; Turner-Stokes et al. 2010).

In the context of spasticity management GAS was shown to provide a sensitive measure of change for a given individual (Ashford and Turner-Stokes 2006; Turner-Stokes et al. 2010). However, limitations have been identified in using the standard GAS approach which relate to a) comparison of scores between individuals or groups and b) data obtained being ordinal rather than interval, undermining the validity of the calculation of the T score (Tennant 2007).

Comparing GAS scores between different individuals or groups is in fact comparing outcome in different constructs in almost every case. However, proponents of GAS would identify the construct as 'goal attainment' or 'achievement of expectation', rather than relating to specific functional items. Nevertheless to carry out GAS group comparison from a mathematical perspective, would require the confirmation of unidimensionality. This leads on to the second issue that scaling is also different in every case, even though the same structure is used and data are therefore at best ordinal (although this is usually not tested). In a simulated evaluation of GAS comparing a linearised, interval scale version of GAS produced by Rasch analysis, with the normal ordinal format, Tennant (2007) found that significant differences occurred with change scores at the extremes of the scale. This means that doubt is cast on the judgement of significant change using ordinal GAS scores, particularly at extremes of the scale. However, this is not entirely unexpected and suggests that less difference is evident in the middle of scale.

Item banking has been proposed as a method of using items identified by GAS for unidimensional constructs within rehabilitation, using item response methods such as Rasch analysis to ensure the interval properties of the data (Tennant 2007). Tennant has suggested that similar goals are often generated within rehabilitation because of the limited options for improvement that are provided by the rehabilitation team. However, patients can have very individual and specific goals for their rehabilitation, which may not always be contained in an item bank. The item-banking approach could be applied,

but would raise the criticism that pre-defined items limit the range of goals and hence the patient centred strength of the approach.

In the area of focal intervention and particularly upper limb function, due to the nature of intervention being focused in an anatomical area, the goals are likely to be even more focused than rehabilitation in general. Therefore, upper limb function may provide an ideal area for construction of an item bank for evaluation of focal spasticity outcome in clinical practice and research. Unfortunately, items to evaluate passive function improvement are still limited and will require further work before an item bank can be considered.

The GAS method currently provides a system of recording goal achievement, which may be used to complement standardised measures in clinical practice. However, for measurement, appropriate standardised measures, with clear psychometric properties will be required (see Chapter 3). Goal attainment scaling provides a possible source of measurement items from goal setting to be included in development of standardised measures or possibly item banks in due course. Items identified may then be compared with existing items from the literature and mapped onto the International Classification of Functioning, Disability and Health (WHO 2002) to confirm the domains being assessed in each case.

1.4 Summary

- Botulinum toxin A reduces upper limb spasticity in patients post stroke, but the improvement in functional ability remains to be established (Elia et al. 2009).
- Possible reasons for the failure to demonstrate functional improvement include:
    a) Lack of appropriate concomitant therapy to maximise the functional improvements (in many studies concomitant interventions are poorly described).
    b) Failure to specify in advance the types of gains (passive or active) that are likely to be achievable – which will vary from patient to patient.
    c) Lack of suitably focused measures to capture the gains that occur (passive and active).
    d) Measuring at the wrong time points. Functional changes may take longer to develop than improvements in spasticity, because they depend on both

reduction of spasticity and improvements resulting from physical interventions.

e) In the case of active function, failure to identify appropriate patients who may benefit.

- Performance measures are applied in the clinic setting and may provide a measure of activity, but currently only record active function and not passive.
- Performance measures also do not reflect how the person undertakes activities in their normal environment and may lack ecological validity.
- Examination of current measures used to evaluate BTX and combined PT intervention may provide possible tools or items as a starting point to evaluate functional outcome.
- Self-report measures are able to record function occurring outside of a test environment in the patient's own setting. While having limitations, they may be useful in evaluating functional (passive and active) outcome following focal upper limb spasticity interventions.
- Goal attainment scaling may provide a useful adjunct to evaluate clinical goal achievement, but can not replace the requirement for psychometrically robust measurement. The analysis of individual goals may provide a possible method for identifying items for inclusion in the development of outcome measures.
- It is essential in the development of tools, that evaluation of the measurement scaling properties of these instruments is undertaken as part of the process.

The studies reported in this thesis address a poorly understood problem; that of the optimal way of measuring function in the upper limb, which represents patients everyday activity, following focal interventions (BTX and PT) for spasticity management. To begin with, current measures used for this purpose are identified and critically appraised and the need for development of a new tool established. Following this, the development of a new measure of active and passive function is described. This development entails a series of sub-studies to generate possible items, develop the measure and then test its psychometric properties. Evaluation of feasibility in the clinical setting is also described and discussed. The following broad hypotheses will be tested during the development and evaluation process.

Hypotheses

1. Measurement items relevant to upper limb active and passive function can be identified from literature and clinical practice sources in a systematic manner.
2. Development of a draft measure from an initial large number of items can be achieved using consultation-based item selection methods with clinicians, patients and carers.
3. Items that capture upper limb function in the context of botulinum toxin and physical interventions for spasticity can form a robust measurement system so that the sub-scale scores can be summed to provide a single summary statistic.
4. The resulting measure can be used to demonstrate improvement in passive function following treatment of upper limb spasticity using BTX and PT intervention.

The objectives associated with these hypotheses are explained in chapter 2.

Chapter 2 Aim and Objectives

2.1 Aim of the research programme

The aim was to identify, modify or develop and then psychometrically evaluate a self-report measure of function suitable for evaluating the impact of spasticity management intervention on the upper limb. Key principles were the assessment of real-life performance and feasibility for use in routine practice following focal spasticity interventions in the hemiparetic upper limb.

2.1.1 Module 1 – Evidence synthesis – literature and clinical practice

Specific objectives (see Hypothesis 1)

1. To identify standardised outcome measures of active and passive function used to assess outcome following focal intervention in the hemiparetic upper limb, which reflect 'real-life' function.

2. To identify candidate items for inclusion in a measure of upper limb function for use following focal spasticity interventions, in the hemiparetic upper limb.

A systematic review was undertaken to determine the nature and scope of current self-report measures to evaluate active and passive function in the hemiparetic upper limb. This served to highlight gaps in existing forms of measurement.

The systematic review was then used to identify existing measures and measurement items that could be included in adaptation or development of a new measure.

In addition, an analysis of goal setting for upper limb spasticity intervention was undertaken to identify additional items for inclusion not present in existing measures. This approach ensured patient and carer involvement at an early stage of measure development and constituted new and original work to underpin the ecological validity of the resulting measure.

2.1.2 Module 2 – Development of the measure

Specific objectives (see Hypothesis 2)

3. To develop a self-report measure to assess both 'active' and 'passive' function in the hemiparetic upper limb following focal rehabilitation interventions.

4. To confirm face and content validity by investigating item relevance for professionals (content), patients and carers (face and content).

Development of the Arm Activity measure (ArmA), a self-report measure of difficulty in arm active and passive function, is undertaken informed by the findings of module one. The structure and scoring is based on measures identified during the systematic review. Item reduction using modified Delphi consultation technique is undertaken, initially involving a group of clinicians, followed by further refinement and confirmation in a much larger group of clinicians and pilot testing by patients and carers. The criteria for the ArmA development are presented in Box 2.1.

Box 2.1 Criteria for measure development

The proposed measure needs to meet the following criteria:

1. Clinical relevance - applicable in the hemiparetic upper limb
2. Include items measuring both:
    a. Active function
    AND
    b. Passive function
3. Assessed in a manner reflective of 'real life' function
4. Practical to apply in everyday clinical practice
5. Valid and reliable for upper limb function evaluation
6. Have established dimensionality of the sub-scales
7. Have established scaling and measurement properties of the sub-scales
8. Responsive to change occurring as a result of intervention

2.1.3 Module 3 – Psychometric evaluation of the measure

Specific objectives (see Hypothesis 3 and 4)

5. To evaluate the reliability, internal consistency, construct validity, unidimensionality and ordinal scaling of the Arm Activity measure (ArmA) – a measure of difficulty in active and passive function.

6. To evaluate the responsiveness and feasibility of the measure when using it to assess outcome following spasticity intervention in the upper limb.

7. To apply the ArmA in measuring change in passive and active function following spasticity management intervention.

Psychometric evaluation of the ArmA also incorporated a cohort study of spasticity management using botulinum toxin and physical interventions in the hemiparetic upper

limb. The main psychometric properties examined were; test-retest reliability, internal consistency, construct validity, responsiveness to change and feasibility.

The cohort study examined the relationship between change in function and spasticity following focal spasticity intervention. Although the main purpose of the study was the psychometric evaluation, the cohort study presented an opportunity to explore this important clinical issue and make a preliminary contribution to knowledge in this area.

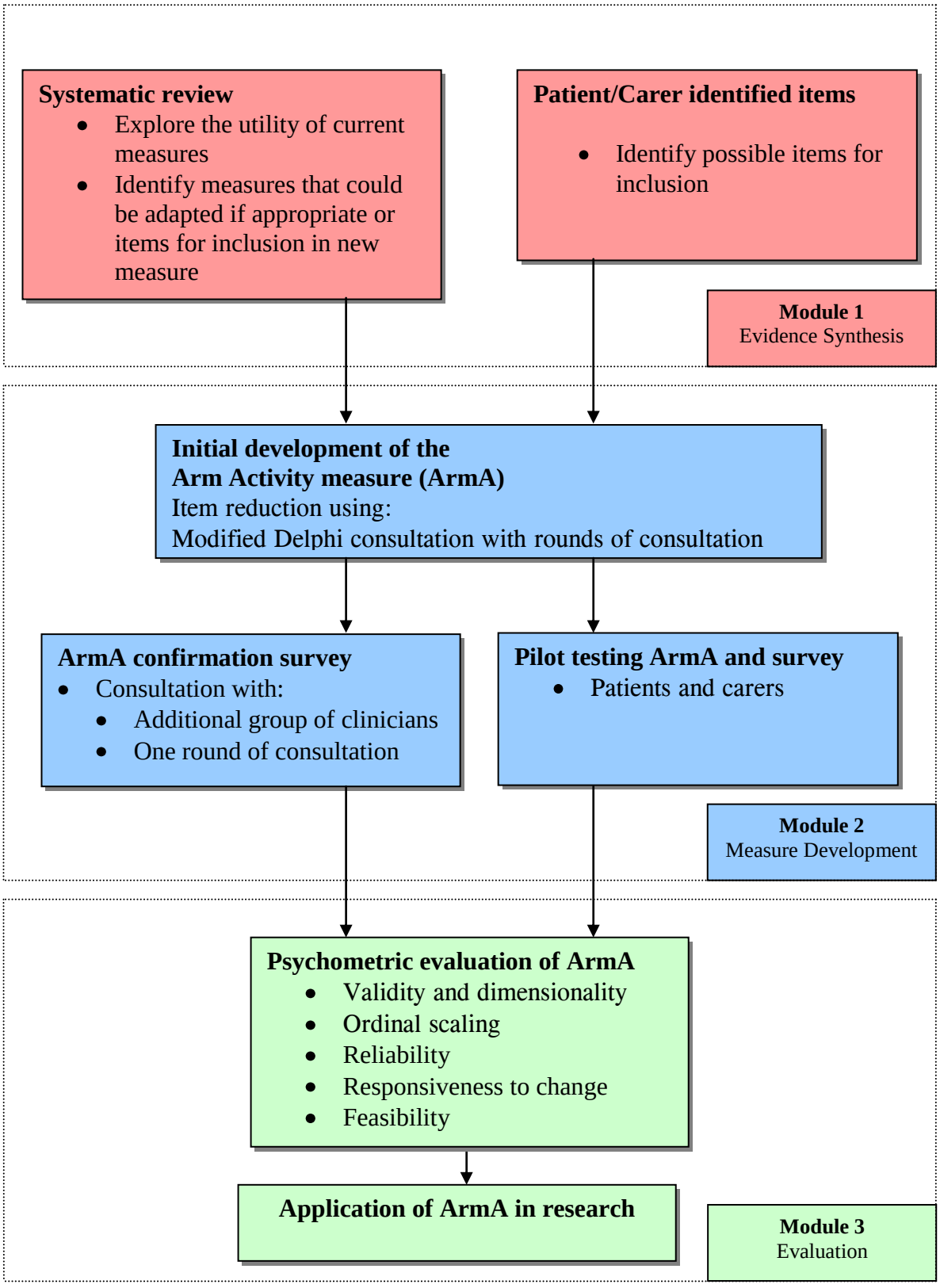
Passive function was addressed, but not active function because change in this sub-scale was expected in only a minority of participants undergoing spasticity management. The utility of ArmA was tested during the cohort study addressing the following hypothesis.

Cohort study null hypothesis: *Improvement in passive function measured by the ArmA at 8 weeks following BTX and PT intervention will not be maintained above baseline levels as the effect of the BTX on spasticity decreases.*

Each module of the thesis builds on the work from the preceding module. Before addressing the need for a new measure of arm function, Chapter 3 presents the theoretical concepts important in the evaluation and development of clinical outcome measures. Chapter 4, through the systematic review, then identifies strengths and limitations of existing measures and highlights the need for a measure of active and passive function in the hemiparetic arm.

Figure 2.1 provides a diagrammatic view of the modules within this thesis.

Figure 2.1 Structure of research programme



Chapter 3 Theoretical issues in measure development

The development of a new psychometric measure typically involves a number of stages and different approaches or strategies may be used in this process. Broadly speaking, a precise conceptualisation of the area to be evaluated is needed followed by consideration of how best to undertake measurement. A process of development is required to develop and then test the tool to ensure that it is actually evaluating the concept (often referred to as 'construct', 'dimension' or 'trait') of interest. If the tool is to be used for measurement, it is essential that it comprises a scale along which changes in the construct can be measured. It is also important that the resulting measure is practical for use; including being feasible for patient and carer completion if self-report is required.

The following chapter will explore what 'measurement' means for constructs such as active and passive function. The approaches to understanding evaluation of such constructs are then discussed with reference to classical test theory (CTT), latent variable methods (LVM) and item response theory (IRT). Methods for evaluating the properties of measurement tools are then considered and related to the approach taken in this thesis.

3.1 Measurement

Measurement is a widely used term that has been defined in a number of different ways depending on the underlying theoretical concept on which it is based. Two broad theoretical approaches can be identified and both of these have further sub-categories. The two theories discussed below are: the 'Classical Theory' and the 'Representational Theory'. Different definitions have also been applied in information theory and quantum mechanics but these are not discussed here due to their limited relevance to this thesis.

3.1.1 The classical definition

The classical definition, which is the accepted standard in the physical sciences, defines measurement as determination or estimation of ratios and quantities (Michell 1990). The 19th century physicist Kelvin gave the following summary:

"when you cannot measure what you are speaking about and express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts advanced to the stage of science" (Thomson 1891).

However this definition of measurement can be traced back much further in history to Aristotle and Euclid (Michell 1990). Aristotle deemed quantity to be one of the fundamental categories of reality and divided it into 'discrete' and 'continuous'. Discrete quantities were defined as categories or natural numbers (e.g. number 2, was thought to be common to all pairs of things), and were essentially descriptive. Continuous quantities were called 'magnitudes' and were intended to designate variables such as length, time or weight.

In the 1920's the classical theory was further clarified by Campbell, who described two types of measurement: fundamental and derived (Campbell 1920). Fundamental measurement requires that the physical world relates directly to physical addition, for example putting 'rods' end to end in a straight line. Therefore, the addition of (in this case) rods in the physical world has similar properties to the addition of numbers in the more abstract condition. Derived measurement is defined by the relationship of a variable, which may not be directly seen or measured to fundamental quantities, which can be measured (e.g. the ratio of mass to volume is a derived measure for a substance's density). This led to the possibility of a more representational approach to measurement.

The concept of 'additivity' is an essential part of the classical theory of measurement. Additivity is seen as the ability to combine numbers (add them together), because they represent addition in the physical world. Additivity is considered so important, because without additivity it is not possible to have ratios of magnitudes and without ratios of magnitudes there is not measurement (Michell 1999). However, Campbell modified the classical theory to allow numbers to 'represent' relationships in the physical world. This resulted in using numbers and the relations between them to represent empirical, non-numerical relations of order and addition.

3.1.2 The representational definition

The representational theory defines measurement as *"the correlation of numbers with entities that are not numbers"* (Nagel 1931). The work of Campbell led to other theorists considering this new perspective on measurement and taking it a step further (Stevens 1946). Stevens defined measurement as:
"the assignment of numerals to objects or events according to rules" (Stevens 1946; Stevens 1951; Stevens 1959)

Stevens identified four principal types of measurement (nominal, ordinal, interval and ratio), although he also acknowledged others such as log-interval scales. These scales were defined in the following manner:

- Nominal (also known as categorical) scales assign items to a category and are a description. For example gender or employment status.
- Ordinal scales place items in rank order, but do not describe the degree of difference between points on the scale. For example, most preferred food to least preferred food. The numbers assigned indicate an order, but do not have true numerical value.
- Interval scales, place items in rank order and have an equal distance between measurement points on the scale, but with no absolute zero. For example, temperature measured in Celsius. Numbers have relative value in this instance, but not necessarily absolute value.
- Ratio scales, meet all the requirements of interval scales, but also have an absolute zero. For example height or mass.

This resulted in Stevens taking representational theory to the point where it allied with a closely related concept of operational theory. In both representational theory, according to Stevens, and in operational theory, *"any precisely specified operation for making consistent numerical assignments to things is measurement"* (Michell 1990). Therefore, measurement is defined by the method (e.g. the questionnaire) rather than an external absolute (e.g. an object with mass in the physical world) as is assumed with the classical definition. Measurement based on Stevens' model became known as 'classical test theory' (CTT), which should not be confused with the classical model of

measurement and therefore in one sense is unhelpful terminology. It is also strange that a model that is comparatively recent is given the 'classical' label and may relate to its wide acceptance as a working definition (Lord and Novack 1968). Methods utilised in Stevens' model will be further explored under classical test theory (page 72).

Stevens definition was widely accepted in psychology from an early stage (Lorge 1951; Green 1954), but has also been widely criticised because of its entirely operational nature (Michell 1999; Borsboom 2005; Borsboom 2006). Michell (1990) has said that in his view; Stevens' definition is mistaken because "*it confuses two quite different practices: measurement (e.g. magnitude in the classical sense) and numerical coding (e.g. discrete quantities in the classical sense)*". Measurement is seen by authors such as Borsboom (2005) and Michell (1990) as specifically related to magnitude representing the real world and should not include the assignment of numbers to categories. However, Stevens' model does provide a system to attempt to organise information from variables that are not so easy to define and evaluate as those in the physical world.

Measurement in common language is often associated with the application of measurement procedures or its operationalisation. However, for such applied procedures to be considered measurement, in the view of some authors, they must relate to an underlying attribute (Kline 2000c). Haig and Borsboom give the following example "*the use of a tape measure does not constitute measurement (of itself); it constitutes the measurement of length, where length functions as a theoretical attribute*" (Haig and Borsboom 2008). Many people, including the general public and many health scientists, do not see the subtle distinction between the attribute being measured and the means of its measurement (e.g. length and tape measure). In the case of physical measurements such as length, the relationship seems clear and does not create any obvious problems. However, in measurement of attributes labelled as 'subjective', the relationship between the measure and the construct measured can be more problematic and difficult to demonstrate.

Variables commonly referred to as 'objective' are often also called 'manifest' and those referred to as 'subjective' are often also called 'latent traits', 'constructs' or 'dimensions'. Measurement of latent variables (e.g. anxiety or quality of life) is more

difficult because it is not possible to measure them directly. They also do not have clear units of measurement in the way most manifest variables do. Although even for manifest variables, the agreement on standardised units of measurement is comparatively recent (e.g. standardization of length and weight measurement was only achieved in the 18th century). In the modern world for example, the manifest variable of height has clear units of measurement that can be used such as centimetres (cm) or inches. The distance between 1 cm and 2 cm and the distance between any subsequent point on the scale and the next is always the same and forms an interval scale. However for latent variables there are usually no accepted universal units of measurement and the units are dependent on which scale is used. The distance between points on the scale will differ (e.g. pain evaluated on Likert or numeric scales). Latent variables are therefore not 'palpable' and are inferred from responses on 'pencil and paper' (or electronic) Likert scales. Measurement of such variables is therefore 'representational' rather than direct.

In the representational theory, the role of numbers in measurement is to represent the real empirical entity (Krantz et al. 1971). Numbers are therefore used to represent empirical relations in the real world. Stevens definition of measurement has therefore been further criticised because it does not address this relationship, but embraces a fully operational approach to measurement (Michell 1990). However, Stevens' approach has strengths in enabling description of simple categories all the way through to ratio scales, but does not demonstrate measurement in terms of magnitude, but rather makes assumptions about it based on the presentation of the data. In practice these assumptions may be reasonable and are certainly practical, which has led to this model being widely applied and accepted (Lord and Novack 1968). However, a problem arises in using numbers related to categories, in anything other than interval or ratio scales. The problem is that if these categories are seen as anything more than labels, in for example nominal scales, then the concept of measurement; using Stevens' own definition no longer makes sense (e.g. categorical or ordinal scales can not be added together). However, Stevens approach to measurement in the form of CTT may provide a starting point for development, but measurement will need to be evaluated by more robustly demonstrating the link between the numbers and the latent variable.

The 'latent variable model' (LVM) has been proposed as an alternative to (and considered by some as an expansion of) classical test theory (Borsboom 2005). The latent variable model attempts to conceptualise theoretical attributes as latent variables. Latent variables are considered the unobserved cause of observed scores (e.g. intelligence and the intelligence questionnaire applied to measure it). The LVM therefore also requires that a statistical model be developed of the relationship between the latent variable and the observed data based on a clear theory. It is only when this relationship is tested and demonstrated, to explain the data seen, that the observation of the latent variable can be interpreted as measurement (Borsboom 2005). The focus of the latent variable model is however targeted on the identification of the construct or latent variable to be measured, but is less concerned with the mechanism of measurement or scaling.

Therefore, within the representational theory of measurement, concepts have been developed to define the relationship between the construct being measured and the scaling of measurements. This approach is also known as 'fundamental measurement' and is concerned with the mathematical relationship of the scale between the variable being measured and the people being measured (Borsboom 2005). The distinction is emphasised between the empirical entities and the numbers assigned to represent them. A system is therefore required to make this relationship clear so that measurement produced can be relied upon to represent the latent variable. Measurement can then be defined as a procedure for identifying values of quantitative variables through their numerical relationships to other values (Michell 1990).

The concept of 'conjoint measurement' was developed to address the requirements of fundamental measurement (Luce and Turkey 1964). Conjoint measurement provides a concept for the identification of quantitative structure other than 'concatenation' (other wise known as physical addition). In fact, it allows a quantitative structure to be identified from ordinal relations to a variable. Therefore, ordinal scales, in some cases developed using a CTT approach, may have methods of conjoint measurement applied (such as Rasch analysis) resulting in interval level measurement scales, which conform to fundamental measurement and additivity as applied in classical measurement. The result of conjoint measurement is that a person's raw score (number of items scored correctly) is a minimal sufficient statistic for their ability (Lord and Novack 1968).

Methods for practically approaching conjoint measurement will be further explored under item response theory in Section 3.2.4 (page 76).

In this thesis a representational perspective to measurement will be applied, defining measurement as "a procedure for identifying values of quantitative variables through their numerical relationships to other values" (Michell 1990). The classical test theory model will initially be applied in ordinal scale development. The latent variable model will then be applied, with consideration of modelling the variables under investigation and initially exploring the ordinal scaling properties of the scale developed. Fundamental measurement in the form of conjoint measurement will not be possible to explore within the scope of this thesis, but will be considered and recommendations made for future work.

3.2 Psychometric methods

Psychometrics is concerned with the theory and techniques for measurement and originated during the 19th century in psychological measure development (Sokal 1971). The primary concern of psychometrics is the measurement of personal attributes and traits (e.g. intelligence), many of which are not possible to measure directly and are therefore considered latent variables. It has two main foci for research; firstly, the development and refinement of theoretical approaches to measurement and secondly, the construction of measures and procedures for measurement (Streiner and Norman 2003; Streiner 2003b).

Psychometrics has a long history in the identification of methods for the development of questionnaires and measures in psychology and education (e.g. intelligence). Its techniques have subsequently been used in the development of many clinical instruments both in rehabilitation and other fields (Hudak et al. 1996; Beaton et al. 2001; Barreca et al. 2005).

Psychometric methods based on CTT are widely used and have been developed and applied in many measures of latent variables (Lord and Novack 1968). Other methods have been developed, which are argued to more directly represent measurement of latent variables without the assumptions inherent in CTT. Item response theory (IRT) is

the general term used to describe these methods, which include approaches such as Rasch analysis and Mokken analysis.

3.2.1 Classical test theory

Classical test theory consists of methods for the development and evaluation of measures and is the underlying basis for the most commonly applied methods in psychometrics (Nunnally 1970; DeVellis 2006). Measures in psychology, education and health care have been developed using CTT and it remains a mainstay of measure development (Hays et al. 2000; Hays et al. 2006).

The origins of CTT began with both the work of Stevens (1946) in defining measurement and also earlier with the work of Edgeworth in proposing that the score for a measure is determined by the real or actual state of the unobserved variable plus error (true score plus error) (Edgeworth 1888). This results in the following equation:

$$\text{Observed} = \text{True} + \text{Error}$$

This model for understanding latent variable scores is inherent in the representative CTT model of measurement and has been used widely in psychology and beyond (Nunnally and Bernstein 1994; DeVellis 2006; Hays et al. 2006). The methodological approach of CTT has been most frequently applied to develop and test measures in psychology (Nunnally and Bernstein 1994). In psychology and fields like rehabilitation there is often a need to seek proxy indicators that provide accurate information about latent variables that are not directly observable (see Sections 3.2.3 and 3.2.4; pages 75-81) (DeVellis 2006).

Within CTT it is accepted that error will occur in measurement between the real value of the variable under investigation and the observed score obtained on the test (DeVellis 2006). These errors are considered random and errors for different items are therefore considered independent of each other, unless testing can demonstrate that this is not the case. Error needs to be minimised in measurement items, to accurately reflect the real value of the unobserved variable. In CTT, considerable effort is expended in selection of items and ensuring that they are valid measures of the unobserved variable.

However, it is still accepted that error will occur and therefore, in general (though not in every case), measures require relatively large numbers of items to account for random error and still produce a reliable proxy of the unobserved variable by reducing the standard error of measurement (DeVellis 2006).

A consequence of managing error by having a large number of items, is that the measurement scale can become very long and can be time consuming to complete, making it less feasible in clinical practice (DeVellis 2006). Another problem in measures with many items is that some items within the scale may be unnecessary to produce a valid overall score, which is often referred to as item redundancy.

There are seven main assumptions that are used to support the use of CTT (Hobart and Cano 2009). Each assumption is very briefly explained as follows:

1. That there is an observed score, a true score and an error score being the difference between the two as discussed above. The true score of the latent variable is by definition unobservable. For a person measured on a particular scale, the true score is assumed constant if measured at different times (in the absence of change). However the error and thus the observed score will vary and result in a range of scores.
2. When a scale is administered to an individual on multiple occasions (in the absence of change), the mean of their observed scores is equal to their true score. This principle assumes that repeated measurements are independent of each other, which will be unlikely when items are repeated for many measures.
3. Errors of measurement are not related to the observed score.
4. Error scores taken from two different scales (of the same construct) are uncorrelated and therefore unrelated.
5. The error score on one scale is uncorrelated with the true score on another scale.
6. Two scales are parallel if, for each person completing the scale, the true score and the error variance are the same.
7. Two scales are equivalent when the true scores for each participating individual are the same except for an additive constant.

In summary, CTT is based on a theoretical assumption that a true score exists, but that the observed score is complicated by the occurrence of error, which is assumed

unrelated to the construct under investigation. Scales are assumed to embody interval level measurement if they fulfil broad criteria such as normal distribution of the study population. Within CTT there is however an acknowledgement that many ordinal scales cannot be assumed to fulfil criteria for interval level measurement because they are clearly ordinal.

3.2.2 Limitations of classical test theory methods

A main criticism levelled at CTT is that the values of the true score and the error score cannot be determined, because they are unobservable theoretical variables (Borsboom 2006; Reeve 2006; Hobart and Cano 2009). This then means that the theory of measurement cannot be tested. CTT therefore does not define mathematically the functional form of the observed score, true score or error score distributions (Borsboom 2006). Hobart and Cano (2009) therefore state that, because the measurement assumptions cannot be tested, they are considered met by most test data sets, which leads to the model being widely applied. However, this may not be justified and Borsboom (2006) argues that modelling of the relationship between the true score (construct under investigation) and the observed score (measure results) should be undertaken as an important component of measure or scale development.

In addition to concerns about the theoretical basis of CTT, issues have been raised about the way in which traditional psychometric methods are used in the evaluation and construction of scales (Borsboom 2006; Reeve 2006; Hobart and Cano 2009).

An important limitation of ordinal data is that, assumptions are needed if the data from multiple items are to be summed into a single statistic. The first assumption, technically made when assigning a 'number' to a point on an ordinal scale, is that in fact the distance between response categories is the same. This is assumed by the fact of assigning a numerical value, but is usually not true. The second assumption, is that total summed scores from such scales are consistent across the range of the continuum represented by the scale, not just as a numerical value but as a true representation of the construct (Hobart and Cano 2009). These assumptions are not evaluated within a CTT approach, and would need to be addressed if fundamental measurement is desired.

A further limitation of using psychometric methods based on CTT to evaluate and develop a scale, is that the performance is dependent on the sample in which it was tested (Reeve 2006). The implication therefore is that scale properties are re-evaluated if the measure is to be applied in a different patient group. Therefore, a problem with CTT based psychometric method is that a person's level of ability cannot be measured independently of the scale used.

3.2.3 The latent variable model

The latent variable model (LVM) can be viewed as an extension of CTT and is therefore discussed briefly here under this umbrella. As discussed earlier, the concept in the LVM is to conceptualise theoretical attributes as latent variables. The latent variables are unobserved (true score), but are measured using the observed variables (observed score i.e. measurement items) (Borsboom 2005). The strength of the LVM over CTT is that it goes beyond an entirely operationalist concept and attempts to relate test scores to an underlying attribute. In the LVM, a formal structure is set up, which relates test scores to the hypothesized attribute. The empirical implications of the model are then deduced and the adequacy of the model is then evaluated by examining the 'goodness-of-fit' with respect to empirical data (Borsboom 2005).

The conceptual framework for LVM originates from the work of Spearman, who developed factor analytic models for continuous data in the context of testing intelligence (Spearman 1904). The work of Spearman in factor analysis was developed further by other authors (Thurstone 1947; Lawley and Maxwell 1963) and led to the conceptual framework of confirmatory factor analysis (Joreskog 1971; Wiley et al. 1973). Factor analysis includes both exploratory and confirmatory factor analyses, which are discussed further in section 3.3.3.3 (page 99).

Factor analysis is one method used to examine the pattern of relationship between different dependent variables (Kline 2000b; Darlington 2010). Kline suggests that although measurement is not ensured by factor structures, these do give useful models about the underlying construct or latent trait (Kline 2000b). While fundamental measurement is of prime importance, it is also important to ensure that the construct or trait of interest is being addressed. Factor analysis may be useful in providing a model

of what construct or dimension scale items are detecting. The limitation of factor analysis is that while it provides valuable information about the construct or dimension is does not address the issue of scaling in the manner of conjoint measurement.

3.2.4 Item Response Theory

In parallel to the factor analytic developments, the concept of latent variable analysis with continuous latent variables was applied to dichotomous data (Guttman 1950; Lord 1952; Rasch 1960; Birnbaum 1968; Mokken 1970). Dichotomous items are those, which only have an either/or response e.g. yes/no or good/bad (i.e. a nominal category), usually with multiple items used in the measurement tool. These measurement approaches, initially used in educational testing, became known as IRT.

The approach of IRT was extended to enable application with 'polytomous' items (Samejima 1969; Bock 1972; Thissen and Steinberg 1984; Molenaar et al. 2000), which corresponds with many scales in psychology and rehabilitation using this type of response option. Polytomous items are those, which have multiple response options for each item.

Borsboom (2006) proposes that one of the main breakthroughs of the past century in psychometric thinking is the realisation that measurement does not consist of finding the right observed score to substitute for a theoretical attribute as in CTT, but of devising a model structure to relate an observed variable to a theoretical attribute as with IRT. Item response theory uses model-based measurement in which latent variable estimates of a state (e.g. level of arm function) depend on both the person's responses and on the properties of the questions that were administered (Embretson and Reise 2000; Hobart et al. 2007). The methods used within IRT are suggested to be superior to CTT by some authors (Reeve 2006; Hobart and Cano 2009), because IRT item parameters are theoretically not dependent on the population sample used in development. They are therefore assumed to be invariant (within a linear transformation) across divergent groups within a population and across different populations (Reeve 2006). However not all authors agree and in practical evaluation, differences have been seen between different patient groups from the same population (Nunnally and Bernstein 1994; Kline 2000b).

The theory of conjoint measurement has been considered in IRT approaches such as Rasch analysis. However, the theory as proposed by Luce and Turkey (1964) has also been applied by other researchers in this area. Michell (1990) has discussed the application of conjoint measurement in Thurstone's theory of comparative judgment in measurement of attitudes (Thurstone 1927). Another example of conjoint measurement is demonstrated in the application of Coombs' theory of unfolding, which explored judgment behaviour among individuals (Coombs 1950). However, although the theory and application of conjoint measurement is not new, it has not been extensively applied in the manner of CTT. This may be due to difficulty in applying and understanding these techniques or limitations of IRT (see the following section), but the need to attempt fundamental measurement is compelling.

In the following section, two methods will be discussed; Rasch analysis and Mokken analysis to further explore the IRT perspective.

3.2.4.1 Rasch Analysis

In this thesis, Rasch analysis will be discussed under the umbrella of IRT. In rehabilitation research the most frequently applied IRT approach has been Rasch analysis (Dickson and Kohler 1996; Penta et al. 1998; Penta et al. 2001; Van de Winckel et al. 2006; Martin et al. 2007; Forkmann et al. 2009). For example, ABILHAND is a measure of upper limb active function, developed by Penta and colleagues (Penta et al. 1998). The measure was developed from a pool of items, which underwent Rasch analysis, to produce the final measure, which using the model has interval level measurement. All items in the scale form a hierarchy, with each measurement item being more difficult than the item before. This produces a measure, which provides consistent and comparable interval data and results in a person's raw score (number of items scored correctly) being a minimal sufficient statistic for their ability (Lord and Novack 1968; Andersen 1977; Perline et al. 1979).

The Rasch model has two main components (Rasch 1960). The first component is that a person's response to an item is governed by two factors: the ability of the person and the difficulty of the item on the construct (Pallant and Tennant 2006). The second

component is the Rasch criteria of invariance – which requires that the relative location of any two persons on the continuum should be independent of the items used to make that comparison (Andersen 1977; Pallant and Tennant 2006). When a set of items is used as a measurement instrument, the aim is to place the items on a continuum onto which people can be measured (Andersen 1977; Tennant and Conaghan 2007b; Hobart and Cano 2009).

Rasch analysis is considered by some authors to be a one parameter model within IRT (de Koning et al. 2002; Raykov and Marcoulides 2011). Two and three parameter models are also available; Rasch analysis is therefore the simplest example of parametric IRT. However advocates of Rasch view it as distinct from other IRT methods because it does not seek to explain the data, but ensures that the data, using the model, are appropriate for interval measurement (Massof 2005; Hobart and Cano 2009). Rasch and other item response models use mathematical constructions to consider the way items and people function related to the latent variable. However, Rasch requires that data fit the model. If they do not fit, then the analysis seeks to identify why, and if required, removes data, re-collects data or re-conceptualises the construct. In contrast, other IRT methods aim to find the item response model that best fits and therefore explains the data.

A key feature of IRT is that a generalised linear relationship is expected between the items and the latent variable, with Rasch analysis being an exception, where a log-linear model is instead expected (Tennant and Conaghan 2007b). The emphasis in the Rasch model is therefore the fit of the data to the model, which produces a form of conjoint measurement corresponding to the concept of magnitude in the classical measurement model.

Hattie showed in a simulation study, that the Rasch model may be inappropriate to assess unidimensionality, because it failed to differentiate between unidimensional and multidimensional items (Hattie 1985). However, the Rasch model assumes a unidimensional model and does not set out to identify unidimensionality as its main aim. Software packages performing Rasch analysis provide an evaluation of principal components for the residuals (after the first principal component has been removed). Thus providing a confirmation of unidimensionality but not as the primary function of

Rasch analysis (Tennant and Conaghan 2007b). Gillespie and colleagues proposed Mokken analysis as one method that can be used to assess unidimensionality in the preliminary stages of scale construction (Gillespie et al. 1987).

3.2.4.2 Mokken analysis

Mokken analysis is often referred to as non-parametric IRT because it evaluates the ordinality of the scale but does not attempt to produce an interval scale. Mokken analysis, as with Rasch, is a probabilistic version of Guttman scaling (Van der Lee et al. 2002; Zwinkels et al. 2004). Guttman scaling requires that responses to an item exactly fit a predetermined model drawn as an item characteristic curve (ICC), items therefore have a fixed hierarchy. In item response theory, an ICC describes the relationship between a latent trait and performance on an individual item. In Rasch and Mokken analysis, the probabilistic version of the Guttman curve allows for small variations from the model, but requires that data largely conform to the model with a predicted fit to the ICC.

Gillespie and colleagues identify one difficulty with Guttman scaling being, that the method assumes a deterministic model with measurement error excluded (Gillespie et al. 1987). Deterministic in this context means that the ICC, which represents the probability of a correct response on the item to the latent trait, must match the functional form dictated by the Guttman model. In practice, Gillespie and colleagues (1987) consider that this poses a difficulty in having no clear criteria for deciding if deviation from the scale represents measurement error in a set of items that are otherwise unidimensional or whether the deviations indicate that, the items lack unidimensionality.

Mokken analysis avoids this problem by proposing a stochastic relationship between the item and the latent trait. The stochastic relationship means that the model allows for variation in the shape of the ICC's functional form. Using the proposed stochastic relationship as a reference starting point, Mokken analysis then provides criteria for deciding if a set of items form a unidimensional scale, and whether a particular item should be included or not. It also allows the addition of items to an existing scale and re-evaluation of the unidimensionality (Gillespie et al. 1987).

The Mokken model treats the latent trait as a single construct along which a person's location can be identified and also an items location or difficulty (van Abswoude et al. 2004). So, provided a unidimensional set of items is identified, the person parameter can be estimated by the number of items to which a person responds positively (referred to as the person score). The item parameter can also be estimated by the proportion of people who respond positively (referred to as the item score).

Mokken analysis therefore evaluates four assumptions.

1) That items form a unidimensional scale.
2) That item scores are locally independent, which means that item scores are independent within a group of persons with the same degree of the construct or trait being measured (Van der Lee et al. 2002).
3) That the 'item response function' for each item is a non-decreasing function of the trait, which means each item would be expected to change in the same direction and to a related degree of change in performance measured against the construct or trait (Van der Lee et al. 2002).
4) That 'item response functions' do not intersect, meaning that the items in the scale have an invariant hierarchical ordering across the latent trait in many different individuals when tested (Van der Lee et al. 2002; Zwinkels et al. 2004). Assumptions (1) to (3) need to be satisfied to accept differentiation between respondents (Monotone Homogeneity), and in addition, assumption four is required to accept differentiation between items (Double Monotonicity). See Section 3.3.3 for further discussion and application of Mokken analysis.

Hobart and Cano (2009) consider the strength of IRT methods, particularly Rasch analysis, to be the articulation of a theory as a mathematical model. Unlike CTT, IRT focuses on the relationship between a person's unobservable score (on the underlying trait) and the probability of responding to one of the response categories on a scale item. They see the application of a mathematical model in this way having three main advantages.

1. Because the model predicts the relationships between variables, it enables evaluation of the consistency of the data to determine how observed data 'fit' the predictions.
2. It enables predictions to be made for the future (e.g. that data collected from another sample will fit the model).
3. The strength of Rasch analysis in particular, comes from the analysis of deviations of the observed data from the predications of the mathematical model. Deviation of items can then be further examined and changes to item categories or removal of items can be considered to allow conformity to the model, which is closer to providing conjoint measurement.

3.2.5 Limitations of Item Response theory methods

Kline (2000), Nunnally and Bernstein (1994) have made criticisms of the Rasch method, which are presented as three main points and also apply to Mokken analysis.

The first criticism is that Rasch analysis assumes that items are equally discriminating, when no test would be so constructed because it makes sense to vary the discriminatory power of items in a test of ability. This is despite Rasch measurement being theoretically 'population free', which means that a measure developed using this method, operates in the same manner irrespective of the population group to which it is applied. However, this criticism does not seem entirely justified, given that items are designed to form a hierarchical scale and therefore if a person is operating low down on the scale they will not be expected to complete the harder items at the top of the scale. A further criticism of Rasch measures being population free is that while in theory this should be true, in practice in rehabilitation this may not be the case. If measures are developed on smaller samples, as is often the case in rehabilitation measures, variation may be seen in the operation of measures. This may be because the different groups are actually different populations as seen in the development of ABILHAND in rheumatoid arthritis and then in stroke (Penta et al. 1998; Penta et al. 2001).

Kline also comments that Rasch analysis also assumes that subjects do not attempt guessing or give socially desirable responses to questions. This is a reasonable

criticism, which has particular relevance for aptitude or educational tests but may not be as significant in measures of health outcome.

Second, large samples need to be tested if reliable population free scaling is to be established (Lord 1974). In addition, in construction of item banks by some authors, it has been identified that virtually no items fit the model if enough calibration is carried out (Kline 2000; Kline 2000b). However for development of measures using both CTT or IRT methods, large samples are often needed and Rasch analysis is probabilistic and designed to accommodate a small degree variation from the model. It is therefore likely that if an exact match to the Guttman scale is being expected, then items will not fit the model.

Third, Kline identified that a problem may occur with the dimensionality of Rasch scales. In a study by Barrett and Kline applying Rasch analysis to the four sub-scales of the Eysenk Personality Questionnaire, a meaningless Rasch based scale was produced with the exclusion of a number of items (Barrett and Kline 1981). However, this does not of itself demonstrate that Rasch is not working to produce a hierarchical scale, because all approaches are dependent on the selection of items, which at least have some relevance to a single construct, rather than reflecting four constructs. Items may seem to fit the model, but this is likely to be coincidental if they have no theoretical relationship to the construct.

3.2.6 Comparison of CTT and IRT methods

Embretson (1996) compared CTT and IRT methods with reference to six measurement rules. Kline (2000) has also made comments comparing CTT and IRT. The comparisons made by Embretson and the comments made on these by Kline have been combined into the following six points.

1. Standard error of measurement. In CTT, the standard error of measurement allows confidence limits to be set, for individual scores, which apply to the whole population. In Rasch measurement however, estimates can be calculated for each level of the latent trait, which are population free. The implication of this is therefore, that although the Rasch method has an advantage in precision, it becomes most relevant in practice at the extremes of the scale.

2. Reliability and test length. In CTT, longer tests with more items have been identified as more reliable in general. In IRT, shorter tests can be more reliable. However, this is only true for IRT measures which have items, selected to fit the level of trait of the subjects. In other instances, Kline (2000) argues that it is still the longer, full version of the test, which is more reliable than the reduced number of items.
3. Parallel tests. In CTT if scores from different groups of items from the same test are compared it is necessary that they are parallel or equated (i.e. the relationship between the items must be established before they can be compared). In IRT, different sets of items measuring the same trait can be highly correlated indicating they are measuring the same underlying construct or dimension and allow comparison between them. However according to Kline (2000), this can be misleading because in real-life data using the Rasch model, the latent trait does not account for all the variance in the items, as it is predicted to do. Kline may be justified in the assumption that the latent trait may not account for all variance and error may be a factor, although this is not really allowed for in a model such as Rasch.
4. Unbiased assessments of item properties. In a simulation by Embretson, as expected, the Rasch equivalent of item difficulties when calculated from two different groups was highly correlated (Embretson 1996). Kline (2000) comments that it is insufficient to demonstrate this in a model alone. However, despite Kline's comment, this strength of the Rasch method in particular and IRT in general is given some support by Embretson's work.
5. Standardisation. In scales developed by CTT the meaning of scores relies on adequate standardisation. However, in theory, meaningful scale scores can be obtained directly from Rasch measures. However, despite the fact that items and individuals can be matched in the Rasch method, population norms are required to give meaning to the scores.
6. Types of scale. In CTT, it is assumed that an interval scale has been produced by selecting items that are normally distributed in a given population. However as already discussed, these measures are then population specific. This is not considered to be the case, by some authors, for a Rasch developed measure, which is assumed to be population free. However, this apparent strength of

Rasch developed measures may not always occur in practice as already discussed.

IRT is based on item characteristic curves, which are probabilistic in the case of Rasch and Mokken, while CTT is a linear model of correlations and factor analysis producing theoretical relationships between variables. Kline also argues that while there are clear differences there are also similarities between IRT and CTT findings (Kline 2000b). He has identified the following evidence for this:

- Roskam demonstrated that factor loadings of the items in a test were good estimates of the slopes in item characteristic curves, with the proviso that the latent trait was normally distributed (Roskam 1985).
- In addition, Nunnally and Bernstein (1994) have argued that there is a very high correlation between Rasch scales and measures produced using CTT methods. Beaton and colleagues (2005) demonstrated this in the development of the Quick DASH, an upper limb function outcome measure. Development resulted in a similar measure using CTT and IRT approaches. However clinical prioritisation of items by clinicians was felt by the authors to lead to a more clinically applicable measure with better face validity, which was the version finalised for use in their work. Clinical prioritisation also resulted in a scale, which retained the psychometric properties of the CTT and IRT developed versions.

Similar findings have been demonstrated in item analysis for other measures using IRT and CTT methods. Pollard and colleagues showed similar results with core items in a measure of impairment, activity and participation in patients following lower limb joint replacement, however some different items were removed with the IRT method (Samejima's Graded Response Model) and the CTT methods (Pollard et al. 2009). Cano and colleagues evaluated the Cervical Dystonia Impact Profile (CDIP-58) using CTT methods to demonstrate that this Rasch developed measure also conformed to CTT psychometric requirements (Cano et al. 2008). The purpose of the analysis was primarily to provide evidence of psychometric properties to meet current United States Food and Drug Administration guidelines for patient reported measures (PROMs) to be used in trials. The CDIP-58 was shown to be robust using CTT and had already been shown to be robust using Rasch analysis. Therefore, CTT and IRT methods may not be

mutually exclusive as demonstrated by these studies. However, it should be conceded that IRT offers stronger evidence of true measurement and can also support evaluation of the unidimensionality of a scale.

The work of Cano and colleagues also highlights a practical issue that outcome evaluation criteria currently focus on CTT psychometric methods in the evaluation of scales for potential use in studies and trials. In addition, current quality criteria such as the COnsensus-based Standards for the selection of health status Measurement INstruments (COSMIN) (Mokkink et al. 2010) and Quality criteria for measurement properties of health status questionnaires (Terwee et al. 2007), use primarily CTT psychometric method criteria to evaluate the measurement properties of scales. The use of IRT methods alone to develop and validate measures may lead to those measures not being considered for use in clinical practice and trials. This may change as IRT methods become more widely used and understood, but currently some demonstration of CTT psychometrics is useful to gain clinical and research acceptance of measures.

The studies of Pollard and colleagues and Cano and colleagues also highlight possible strengths in using CTT and IRT in combination. This approach has been seen in some measures such as the Impact of Psoriasis Questionnaire, which was developed using CTT methods and then refined using confirmatory factor analysis to more clearly define its sub-scales (Nijsten et al. 2006). Each sub-scale then underwent Rasch analysis to produce a robust measure. Forkmann and colleagues have used a similar approach in developing a depression screening measure (Forkmann et al. 2009b).

Although IRT approaches and Rasch scales in particular are examples of "additive conjoint measurement" (in other words interval scaling), the absence of a unit of measurement that corresponds to something tangible in the real world, raises some of the same concerns as for CTT developed measures (Kline 2000c). This is a fundamental concern, over and above, any practical criticisms to be levelled at application of approaches such as Rasch analysis. However, Kline (2000) concludes that while the approach of CTT has been valuable, its assumptions do not replace the need to strive for 'true' measurement. He also concedes that in the absence of such 'true' measures, representative approaches may still be valuable.

Limitations to CTT and IRT approaches have been identified. Application of CTT approaches are still relevant in the initial development of measurement tools and in particular can assist in the modelling of constructs or dimensions essential for measurement. However, IRT approaches, particularly Mokken and Rasch methods, represent advances in scaling latent variables. As discussed, IRT approaches also have limitations, which should be acknowledged and cannot resolve completely the difficulties of latent variable measurement, but currently represent the best options for scaling of latent variables. Clinical utility is also important in the development of measures and will be discussed in the following section related to clinimetrics and psychometrics.

3.2.7 Psychometric and clinimetric challenges

The following section will briefly discuss the theoretical difference in perspective between psychometrics and clinimetrics and the relation of these concepts to measure utility. Clinimetrics, in common with psychometrics, is also concerned with evaluating measurement properties, borrowing many of its methods from the psychometric literature. However, the approach in clinimetrics is to emphasise strongly the clinical application of the resulting measure (i.e. utility) as a factor in measure development in addition to ensuring robust measurement properties. These theoretical concepts are of relevance to this thesis because clinical utility of the measure is of prime importance in conjunction with robust measurement properties.

A current perceived controversy in clinical measure development exists between the concepts of psychometrics and clinimetrics. The controversy emphasises on one hand the need to ensure robust measurement properties; while on the other ensuring that items of clinical importance are included in measures.

Clinimetrics

Feinstein introduced the term clinimetrics in 1983 and defines it as referring to "arbitrary ratings, scales, indexes, instruments or other expressions that have been created as measurements for clinical phenomena that cannot be measured in the customary dimensions of laboratory data" (Feinstein 1983; Feinstein 1987). Wright and

Feinstein (Wright and Feinstein 1992) state that measurement depends on four components:

1. The phenomenon chosen as the focus of attention
2. A strategy for constructing the measure
3. The single or multiple items that describe the selected attributes of the phenomenon
4. An index that expresses the final rating of the aggregated items

Feinstein suggests the importance of evaluating measurement instruments for validity and reliability has been lacking in the development of some measures in clinical practice. Feinstein also emphasises however, the importance of what he terms *sensibility*, referring to whether the measure appears to measure the intended construct or has face validity.

3.2.7.1 Psychometrics and clinimetrics

The difference between the two concepts has been characterised as psychometrics attempting to measure a single dimension or construct with multiple items using validation methods to demonstrate that all items measure a single attribute. While clinimetrics, measures multiple attributes with a single index and strategies for measure development focus on selecting from a clinical perspective the most important items to be included (Fayers and Machin 2007).

The purpose of both psychometrics and clinimetrics in theory is the production of valid and reliable measures. However clinimetrics is thought to be different from psychometrics in being more concerned with *sensibility* or face validity of a measure while psychometrics gives more importance to the way in which the measurement items (a) correlate with each other and (b) measure the same construct (Wright and Feinstein 1992). Clinimetrics as proposed by Feinstein (1987) appears to advocate the acceptance of clinical measures that combine scores from items from different traits or dimensions, which leads to significant criticism from measurement theorists. A number of clinimetric type scales have been developed, such as the Glasgow Coma Scale (Teasdale and Jennett 1974) for assessment of consciousness and the APGAR scale (Apgar 1953) for assessment of responsiveness of infants. Both of these scales are

multidimensional and are often given single scores, yet provide useful clinical information for decision making. However, they do not conform to the concept of conjoint measurement and it is not meaningful from a measurement perspective to produce total scores for these scales.

Feinstein suggests that a conflict occurs between *sensibility* and standardisation of items to form single constructs. That while clinimetrics may be lacking in standardisation of measurement items and scales, psychometrics may lack sensibility in application of items and measures to clinical outcome (Feinstein 1987). It has been suggested that the two fields can learn from each other and that in particular applying psychometric techniques can progress clinimetrics and the development of measurement scales in clinical practice (Wright and Feinstein 1992).

Clinimetrics has been criticised for being an unnecessary term, which creates confusion in the literature (Streiner 2003b). Streiner states that Feinstein proposed that clinimetrics is a "sub-set of clinical epidemiology", while he states it is a subset of psychometrics. Streiner has also been critical of the suggestion by Feinstein that "all questionnaires in psychology are unidimensional and all those in medicine are heterogeneous" (Streiner 2003b). He suggests that this overlooks the diversity of instruments in both fields and that at the extremes in both areas this may be true but is not for the majority of measures in either field.

In summary, the debate regarding clinimetrics and psychometrics has at times become unnecessarily polarised, which seems counterproductive in meeting the objective of providing clinically useful measures. While the introduction of clinimetrics is useful in highlighting the clinical utility and feasibility of measures, it risks detracting from the requirements for fundamental measurement. Both clinical utility and robust measurement properties are required in useful clinical measurement tools. The following section presents the principles of measure development and psychometric testing used in this thesis.

3.3 Principles of outcome measure development

Donabedian defined health outcome as "a change as a result of antecedent healthcare" (Donabedian 1980). This definition is directly applicable to active function outcomes, but is too narrow when considering passive function outcomes. Passive function outcomes may include the prevention of deterioration rather than just improvement, which is harder to assess. Therefore, health outcome could be more usefully defined as 'achievement of an intended health goal as a result of antecedent healthcare' which is the definition used in this thesis.

Measures may be developed from different perspectives, that of the population or that of the individual. Measures, which are useful in describing a population, are described as '**nomothetic**'. A nomothetic measure is designed to quantify the observation made with population norms (Merriam-Webster Online Dictionary 2009). An example of a nomothetic measure would be the Jebsen Test of Hand Function (Jones 1990), which compares the performance of the individual patient against population norms for patients with stroke and therefore needs to conform to the standards for measurement discussed at the beginning of the chapter.

Measures that are developed from the perspective of the individual are described as '**idiographic**'. An idiographic measure compares an observation with a standard set by the patient and therefore can only be compared with that patient directly (Merriam-Webster Online Dictionary 2009). An idiographic example would be Goal Attainment Scaling (Kiresuk et al. 1994), which allows the patient to set a goal and then compare the outcome of intervention against the goal they have set. This type of approach presents theoretical difficulties however if arithmetic evaluation and comparison against other individuals is undertaken who have goals set in differing goal areas.

Both idiographic and nomothetic measures have strengths and weaknesses. Idiographic measures have the advantage of being responsive and specific to the aims of the intervention, but are more limited in terms of comparisons between individuals and across populations. Nomothetic measures may allow comparison easily between individuals but may not be sensitive to the specific needs of the individual and therefore may not be responsive to change in that individual.

3.3.1 Criteria for a good measure

The development of any measure must include evaluation of its measurement properties (McDowell and Newell 1987).

- A measure must be **valid** in recording what it purports to measure and relevant to the effects of the intervention being examined.
- It must be **reliable** in distinguishing between different patients and must be able to distinguish important change rather than measurement error (Terwee et al. 2007).
- It must be **responsive** to change that occurs as the result of intervention (Wade 1992a).
- For a measure to be used in normal practice it must also be **feasible**, that is simple and practical enough to apply in routine clinical practice (Slade et al. 1999).

(Also, see Glossary)

The Scientific Advisory Committee (SAC) of the Medical Outcomes Trust have proposed criteria for the evaluation of measures regarding their psychometric properties (Medical Outcomes Trust 2002). The SAC criteria contain eight areas to be used for evaluation of measures.

1) Content validity
2) Internal consistency
3) Criterion validity
4) Construct validity
5) Reproducibility
    - Agreement
    - Reliability
6) Responsiveness
7) Floor and ceiling effects
8) Interpretability

Terwee and colleagues (Terwee et al. 2007) have more recently refined and updated these criteria by developing quality requirements for the identified properties of a

measure using CTT. These criteria or criteria of a similar nature based on CTT, are widely used in evaluating the psychometric properties of health related outcome measures (Fitzpatrick et al. 1998; Moe-Nilssen et al. 2008; Scarpelli et al. 2008; Tamber et al. 2009). The majority of psychometric evaluation criteria identified in a recent search by the author are based on CTT as indicated in those above. One exception was from the patient reported outcome measures group, which in addition to CTT approaches to psychometrics also includes reference to the use of IRT under the heading of precision (Fitzpatrick et al. 1998). The inclusion of IRT approaches in this way is welcome, but may be viewed by some proponents of IRT, as not giving sufficient importance to these methods. However including both CTT and IRT methods has strengths in both a) ensuring maximum understanding and acceptance of the psychometric properties of a resulting measure and b) ensuring robust measurement properties.

3.3.2 Item Generation

The first step in development of a measure is usually to identify items for inclusion (Streiner and Norman 2003). Items may be taken from a number of different sources:

A) Critical analysis of the published literature (Hobart et al. 2001). At this stage a number of possible items may be identified from pre existing scales (Hobart et al. 2001).

B) Analysis and synthesis of the views of patients, carers and clinicians (Streiner 2003b), which may also be used to confirm items identified from literature sources or initially generate new items (Hicks 1999; Hobart et al. 2001).

C) Analysis of clinical practice also provides possible items to address areas that have not been included in previous measures (Streiner 2003b).

A) Systematic reviews may be of value in identifying relevant measures already developed as well as additional items. This approach has been used in measure development in neurological rehabilitation as well as other areas of rehabilitation management (Bot et al. 2004; Terwee et al. 2006; Rowland and Gustafsson 2008; van de Ven-Stevens et al. 2009). This method can produce a large pool of items, but risks reflecting the accepted wisdom on a subject and in particular may be prone to reflect the views of clinicians rather than patients and carers.

B) Views of patients and carers may be more relevant to the real-life context, which may be of prime interest in evaluating functional outcome. Items identified by patients can also be "more specific and concrete" than those generated by clinicians and are important to consider for this reason (Lomas et al. 1987). In addition, they are likely to add to face and construct validity in the final measure.

Obtaining the views of patients and carers may take a number of different forms. Consultation with potential users (patients and carers in this instance) may be achieved by applying methods such as focus groups or semi-structured one-to-one interviews. Both focus groups and semi-structured interviews allow participants to share their views and enable incorporation of participant's experiences of healthcare and difficulty in functional performance into measure development. However, group discussions may have advantages over semi-structured interviews for identification of measurement items. While semi-structured interviews tend to be directed by the researcher (Reed and Roskell-Payton 1997; Smith et al. 1997), group discussions lend themselves to direction by the participants with researchers taking a more passive role (Bender and McKenna 1994). In addition Sim and colleagues have suggested that group interaction maybe more spontaneous and Kitzinger has suggested that focus groups allow participants to vocalise previously unarticulated thoughts and reveal common views with others in the group (Kitzinger 1994; Sim and Snell 1996).

While there are advantages to focus groups, there are also potential problems that may occur with group interaction. The involvement of group members may be limited by personal inhibitions, competition to speak or get their particular point across by group members and the social convention of turn-taking in conversation.

C) Goal setting in rehabilitation practice provides a unique opportunity for interrogation of clinical records. However, analysis of clinical practice may take a number of forms; including reviewing clinical records, clinical team decision-making and evaluation of clinical results. The specific review of rehabilitation goals provides an innovative method of obtaining patient and carer views. In rehabilitation the use of goals set for a specific intervention, may provide a potential 'pool' of items to include in the development of measurement tools.

The advantage of this method is that it focuses on the particular area under investigation, avoids the problems associated with group interactions and elicits the patients (and sometime carers) perspectives on what is important to them. However disadvantages may include the researcher or clinician influencing the patient in the goal setting process (Reed and Roskell-Payton 1997), lack of reflection on the whole patient group (goals may be specific to that individual) and artificial restriction of goals to a very specific area. Nevertheless using rehabilitation goals as a source of possible items, is both efficient at identifying the issues important to patients and carers and also ensures face validity of the items.

In a measure of arm function, items may include many diverse activities. It is not possible to include all possible items in a measure, because this would make the resulting scale long and time consuming to complete (Messick 1980; Streiner 2003b). The items initially generated will therefore, in due course, need to be reduced to make the scale feasible for clinical application.

3.3.2.1 Item reduction

Once the possible items have been identified, it is necessary to select the 'best' items to be used in the measure. Item reduction may take different forms, which broadly divide into two categories; statistical methods to differentiate the items from a measurement perspective once data has been collected, and more intuitive approaches using the opinions of experts to identify the most important items to reflect the construct, for example Delphi Consultation (Reid 1988). Delphi consultation, while still enabling interaction with multiple individuals, avoids some of the problems identified with focus groups such as personal inhibitions, competition to speak and turn-taking in conversation (Burns et al. 2003; Finger et al. 2006).

Experts used in approaches such as Delphi consultation may be patients, carers or clinicians. The priority for removal of items will firstly be those that are least relevant to the construct being measured and secondly those that are redundant or unreliable. The process ensures that items, which are clinically important, relevant to the construct and enhance utility and feasibility are included. However, the Delphi approach does not ensure the measurement properties of the resulting items and methods to evaluate this

will therefore be needed. Once items for possible inclusion in a measure have been identified, it is necessary to consider how responses to these items will be recorded, measurement properties and the possibilities for analysis that this presents (McDowell and Newell 1987).

3.3.2.2 Relationship of measurement to the construct or trait

Methods to evaluate the relationship of items to the latent trait or construct, such as exploratory factor analysis in CTT and methods such as Mokken in IRT, may result in a reduced number of items. In the case of exploratory factor analysis, the focus is to determine the constructs covered by the items, but this may result in the exclusion of items which do not fit any construct or form a construct which is not the object of investigation (Kline 2000b).

In IRT methods, such as Mokken analysis and Rasch analysis, items may be removed because they do not fit the scale, but again the primary aim is not the removal of items. In the case of Mokken analysis, the aim will be to produce a unidimensional ordinal scale.

3.3.3 Psychometric properties of measures

The following section explores the psychometric properties that apply to measures of function in the upper limb and discusses the benefits and challenges of considering these issues during measure development. The approach taken is primarily one of initial evaluation using CTT methods with addition of preliminary evaluation using IRT methods in the form of Mokken analysis. The criteria used to evaluate psychometric properties are primarily based on those by Terwee and colleagues (Terwee et al. 2007) incorporating the work of the Scientific Advisory Committee (SAC) of the Medical Outcomes Trust (Medical Outcomes Trust 2002). In addition, IRT methods of scale evaluation for measurement have also been considered as advocated by Hobart and Cano (2009), but using Mokken analysis to assess for ordinality of the scale as an initial step in IRT evaluation (See Table 3.1).

Table 3.1 Quality criteria used in this thesis adapted from Terwee and colleagues (2007).

Attribute	Criteria	Evaluation
Validity		
	The degree to which the instrument measures what it purports to measure	
	Face	**Confirmation that the measure is appropriate to its intended use** E.g. Feedback from the target population that the items of interest are comprehensively represented
	Content	**The concept to be measured, and the intended target population (who should be involved in scale development)** Give a clear description of the structure of the scale and intended level of measurement Describe empirical basis for selection of item content and combination (Positive rating if the above are clearly described and target population was involved in item selection)
	Criterion-related	**Evidence that the scores of the measure are related to an accepted gold standard** (Give rationale for choice of criterion measure) (Positive rating, if a convincing argument is presented for the gold standard and correlation is at least 0.70)
	Construct-related	**Evidence that the scores relate to other measures consistent with theoretically-driven hypotheses** The extent to which the scores correlate with other measures of similar concepts (convergent validity) and do not correlate with unrelated measures (divergent validity) (Positive rating: hypotheses formulated and at least 75% of results are in accordance with these)
	Unidimensionality	**Information on unidimensionality and rationale for deriving scale scores (measurement)** Describe any intended subscales built into the conceptual design Exploratory and Confirmatory Factor Analyses within target population – how many factors – what does each represent? (Positive rating if factor analysis performed on adequate sample size (n= (7 x no of items) and at least 100) Mokken Analysis using a monotone homogeneity model to confirm dimensionality and preliminary measurement properties in the scale (minimum values of H indicating a strong scale is 0.5, a medium scale 0.4 and a weak scale 0.3).
Scaling		**Confirmation that items form a hierarchical scale conforming to either ordinal or interval scaling.** Mokken Analysis applying a monotone homogeneity model, double monotonicity model or conformity to Rasch measurement.

Internal consistency		**The precision of the scale based on interrelatedness of the scale's items** Item-total correlations (correlation of each item with the total score excluding that item) Cronbach's alpha calculated for each subscale – Positive rating: alpha should be between 0.70 and 0.95 (higher than this suggests item redundancy)
Reproducibility	***The degree to which the measure is free from random error***	
	Agreement	**Extent to which scores on repeated measures are close to each other** Minimal Important Change (MIC) < Smallest Detectable Change (SDC) outside the limits of agreement OR arguments that agreement is acceptable
	Reliability	**Stability of the test over time (repeatability, intra-rater reliability) and between different observers (inter-rater reliability)** (Time period should be long enough to prevent recall, but short enough that no clinical change has occurred) Total scores – intra-class correlations (ICC) or Bland and Altman Limits of agreement for continuous (interval) data Item-by item analysis of agreement using weighted kappa statistics for ordinal data (Positive rating: ICC and kappa coefficients should be at least 0.70 in a sample of at least 50 patients)
Responsiveness	***Ability to detect change over time where real changes occur***	
	Change	**Evidence of change in longitudinal analysis** – significant differences (Wilcoxon z or paired T tests), Effect size estimation Ability to distinguish clinically important change from measurement error Positive rating, the (SDC)* or the Limits of Agreement should be smaller than MIC*; or Gyatt's Responsiveness Ratio: MIC/SDC = at least 1.96 (NB Excluding Wilcoxon, these are parametric tests suitable really only for continuous normally distributed data) Alternatively, the area under the Receiver Operating Characteristics (ROC) curve should be at least 0.70

Interpretability		
	The degree to which easily understood meaning can be assigned to the quantitative scores	
	Clinical meaning	**Describe how the tool should be reported and interpreted** e.g. sub-scores, total scores etc Provide information about what change in score would be clinically meaningful – define MIC – "the smallest difference which patients would perceive as beneficial and would mandate a change in the patient's management (in the absence of troublesome side effects and excessive cost)". Comparison of change in groups expected to change with those not (e.g. responders vs. non-responders, active treatment vs. placebo) (Positive rating: descriptive statistics given for at least 4 relevant subgroups of patients and MIC defined)
Floor/ceiling effects		**Floor or ceiling effects are present when >15% of subjects achieve the highest or lowest possible score.** (Positive rating: no floor or ceiling effects in a sample size of at least 50 subjects in the target population)
Burden		
	The time, effort or other demands of administering the instrument	
	Time to administer	Information on average and range of time taken to complete the instrument Any special requirements or restrictions – e.g. training, level of professional expertise to apply it Under what circumstances is it not suitable? (Positive rating: if burden is described and acceptable to target users and clinicians)
Alternative modes of administration		Describe if these are available Evidence of reliability, validity, responsiveness, interpretability and burden for each administration Information on comparability of the alternative modes
Cultural and language adaptations		Describe if these are available Methods to achieve conceptual and linguistic equivalence. Any significant differences between the original and translated versions – how any differences were reconciled

*Smallest Detectable Change (SDC), Standard Error of Measurement (SEM), Minimal Important Change (MIC)

3.3.3.1 Criteria for psychometric assessment

The criteria of Terwee and colleagues were used in a systematic review of upper limb function measures for rehabilitation in patients after musculoskeletal problems (Bot et al. 2004). In addition to the original criteria, feasibility has been added and is discussed below. Construct validity has been expanded, with the inclusion of unidimensionality and evaluation of ordinal scaling using Mokken analysis (see Table 3.1).

Psychometric terms used in this thesis are referred to in Table 3.1 and are included for reference in the glossary.

3.3.3.2 Validity

Face and content validity should be established before a scale is used to evaluate the outcome of clinical intervention and are often considered in the development of a measure (Streiner and Norman 2003). **Face validity** is important because:

1. It increases cooperation and motivation among respondents
2. It attracts respondents
3. It reduces dissatisfaction among respondents
4. Makes it more likely that policy-makers and funders will accept findings

(Nevo 1985)

Face validity is usually established by asking a panel of users to review the measure, but can also be addressed by involving users in the process of development.

A closely related concept to face validity is **content validity**, which is similar, but evaluates that the instrument covers all the relevant concepts or domains (Streiner and Norman 2003). A tension often exists between ensuring that a measure contains all the relevant items related to the construct being measured and yet is still short enough to be feasible to apply in clinical practice (Tansella and Thornicroft 2001).

Criterion validity is usually the comparison of the new measure with an existing gold-standard measure both applied at the same time (**concurrent validity**) or comparing the predicted result with the actual outcome (**predictive validity**) (Streiner and Norman

2003; Barreca et al. 2005). In many cases, the reason for developing a new measure is that a comparable scale does not exist and therefore comparison with a gold standard is not possible.

Evaluation of **'construct validity'** may be undertaken by linking the detection of change identified by the new measure to a prediction or hypothesis of how change should occur (Cronbach and Meehl 1955; Streiner and Norman 2003). The term 'construct validity' therefore originates from the idea that the new measure evaluates the construct it has been designed to measure (Cronbach and Meehl 1955), however the term is often applied even more broadly in current practice. It may be possible in some cases to link change in an existing measure, related to the construct, with change in the new measure and thus demonstrate 'construct validity' (Turner-Stokes et al. 1999).

3.3.3.3 Unidimensionality of sub-scales

Measures may be developed to evaluate a single construct or they may contain a number of sub-scales addressing different constructs, which have a relationship to overall function. Multi-dimensional measures are common in evaluation of function and may contain sub-scales evaluating different, functional areas or constructs (McPherson et al. 1997). However, for conjoint measurement, evaluation needs to be focused on a single dimension and therefore sub-scales will often be used for multidimensional measures. Attempts have been made to apply conjoint measurement to multidimensional measures using the theory of multidimensional scaling (Michell 1990). However, such methods have come under considerable criticism (Wiener-Ehrlich 1978; Tversky and Gati 1982), because the dimensions of such scales are unclear and therefore measurement is questionable. Therefore, for measurement to be possible using current methods, a unidimensional scale is essential. To establish unidimensionality (or constructs) of scales or sub-scales CTT approaches such as factor analysis may be used or IRT approaches such as Mokken analysis.

Factor Analysis

Factor analysis is used to examine the pattern of relationship between different variables (Kline 2000b; Darlington 2010). Kline suggests that factor structures, give useful

models about the underlying construct or latent traits, which relate to a single dimension (Kline 1994; Kline 2000b).

The term factor analysis (FA) includes a number of different methods used to examine how underlying constructs influence responses on measurement items (or other variables dependent on what is being investigated) (Kline 1994; Raykov and Marcoulides 2011). Two broad types of FA are possible: exploratory factor analysis and confirmatory factor analysis. Exploratory factor analysis (EFA) attempts to discover the constructs influencing a set of responses (DeCoster 1998). Confirmatory factor analysis (CFA) evaluates if a specified set of constructs (usually identified through EFA) is influencing responses in a predicted way (DeCoster 1998).

The model underlying FA proposes that each item response is influenced by underlying factors, which are common to all items or groups of items. In addition, each item also has a unique underlying factor. Factor analysis is then undertaken by examining the pattern of correlations (or covariance's) between the observed items (DeCoster 1998). Items which are highly correlated (positively or negatively) will usually be influenced by the same factors. In contrast, those that are relatively uncorrelated are likely to be affected by different factors.

Principal Components Analysis

The aim of principal components analysis (PCA) is conceptually slightly different to FA. Exploratory factor analysis and PCA are often considered to be synonymous, however some differences exist. The concept underlying PCA is to reduce the dimensionality of a data set consisting of a large number of interrelated items (Jolliffe 2002). The PCA method operates by transforming the interrelated items to a new set of variables (the principal components), which are ordered so that the first few components retain most of the variation present in all the original variables (Jolliffe 2002).

The differences between EFA and PCA originate from the different models on which they are based. With EFA the responses to items are assumed to be based on underlying factors, while in PCA the principal components are based on the measure's responses (DeCoster 1998). The PCA method has been frequently applied in rehabilitation research to identify the number of dimensions underlying a set of interrelated items

(Bedard et al. 2001; O'Rourke and Tuokko 2003). PCA has been used for evaluation of measurement instruments, to identify the principal components contained within the item responses, before evaluation of scaling properties. In addition, a combination of PCA and IRT approaches has been used to evaluate and develop measures.

An example of this is provided by Siegert and colleagues who undertook a 5-step analysis of the factor structure and dimensionality of the Zarit Burden Interview in a sample of 222 carers (Siegert et al. 2010).

- Steps 1 and 2 involved a CFA (using AMOS 16), followed by Rasch analysis of the full scale.
- Step 3 used a PCA with rotation of the resulting factors (Siegert et al. 2010).
- Steps 4 and 5 included further Rasch analysis of the two sub-scales identified by PCA and was followed by further CFA.

This demonstrates a mixture of methods used in a structured manner to clarify the measurement properties of the scale. Other authors have also used the combination of both CTT (including PCA) and IRT methods in the development and evaluation of clinical measures (Smith et al. 2007; Cano et al. 2008), representing a robust approach to development.

Results from principal component analysis can be evaluated by initially considering Eigenvalues above 1 according to the criteria by Kaiser (Kaiser 1960). Kaiser considers Eigenvalues above 1 to indicate the principal components of value and provides an objective criterion for retention of components (Teo and Chong 2006). Following evaluation of Eigenvalues, Scree plots can then be examined to confirm the findings. Consideration of Eigenvalues and Scree plots in identification of principle components is considered to be more robust by some authors than using Eigenvalues alone (Butler et al. 2006). To confirm these findings and to provide more objective criteria for the acceptance of the components identified, a Monte Carlo analysis may be carried out according to the method by Horn (Horn 1965). The Monte Carlo analysis involves parallel analysis of findings against predicted objective criteria to determine acceptance of the principal component findings.

Sample size for PCA

Sample size is important for statistical procedures such as PCA or EFA, because smaller samples may lead to erroneous conclusions (Osborne and Costello 2004). These procedures optimise the fit of the model to the data, but no sample is able to perfectly reflect the population. Over-fitting can also occur and result in key errors such as extraction of erroneous factors or mis-assignment of items to factors (Tabachnick and Fidell 2001).

Sample size for PCA therefore remains the subject of confusion and considerable debate. Approaches to sample size also differ with some advocating estimates based on total sample and others on ratios of numbers of subjects to scale items. In general, samples of a larger size are preferable to smaller samples, because they tend to minimise the probability of errors, maximise the accuracy of population estimates, and increase the generalised application of results (Guadagnoli and Velicer 1988). Unfortunately, many of the sample size recommendations for PCA and EFA are limited and often have minimal empirical evidence (Guadagnoli and Velicer 1988). Further discussion of sample size for PCA is undertaken in Chapter 7 in relation to psychometric evaluation (see Section 7.3; page 191).

Different authors have suggested a range of sample sizes for PCA. Nunnally and Bernstein recommend a ratio of 10:1 subjects to items (Nunnally and Bernstein 1994) and Terwee and colleagues (2007) a ratio of 7:1 with a total sample of 100 or greater. Norman and Streiner have discussed sample sizes between 50 and 1000 (Norman and Streiner 2000). A number of studies have used multiples of 100, but all below 1000 for PCA in rehabilitation research (Smith et al. 2007; Cano et al. 2008; Siegert et al. 2010).

3.3.3.4 Scaling

Measurement principles are discussed at the beginning of this chapter (Section 3.1; page 65); the following section will focus on how Mokken analysis can be applied for evaluation of ordinal scaling properties. Mokken analysis is considered in more detail, due to the preliminary nature of the evaluation in this thesis, the relatively small number of participants (n=92) and the desire to establish ordinality in the scale as appose to Rasch analysis, which operates on transformed scores. Confirming the ordinality of the

measure has particular relevance for a clinical tool where raw scores may be used to interpret patient outcome.

In Mokken analysis, the 'item response function' for each item is a non-decreasing function of the trait. Meaning that each item would be expected to change in the same direction and to a related extent, to change in the construct or trait (Van der Lee et al. 2002). The Mokken model treats the latent trait as a single construct along which a person's location can be identified and also an items location or difficulty (van Abswoude et al. 2004). The final assumption evaluated in Mokken analysis for Double Monotonicity, is that items in the scale have invariant hierarchical ordering across the latent trait in different individuals when tested (Van der Lee et al. 2002; Zwinkels et al. 2004).

Monotone Homogeneity

As discussed (Section 3.2.4; page 76), the Mokken model specifies the relationship between the item and the latent trait using an item characteristic curve. An important feature of the Mokken approach, which is different to other IRT models, is that it makes no assumptions about the functional form of the item characteristic curve. This is the reason for Mokken analysis being referred to as non-parametric.

Double Monotonicity

Double Monotonicity, in addition to the assumptions of Monotone Homogeneity, assumes that the item response curves do not intersect. So that, the probability of a positive response decreases with the difficulty of the item, which will still allow for variation in the shape of the ICC, but will reflect the greater difficulty from the previous item through the entire range of that item on the latent trait (Gillespie et al. 1987).

The Double Monotonicity model is a 'special case' of the Monotone Homogeneity model, the fit of the Monotone Homogeneity model should be explored before non-intersection of item response functions are investigated (Molenaar et al. 2000). Monotone Homogeneity was therefore considered in this thesis before further evaluation of the scale developed.

Application of Mokken Analysis

Mokken analysis produces the H coefficient of scalability (Loevinger 1948), which is a measure of the accuracy of ordering participants. The minimum values of H indicating a strong scale is 0.5, a medium scale 0.4 and a weak scale 0.3 (Roorda et al. 1996; Van der Lee et al. 2002; van Schuur 2003). van Schuur also identifies that no item in a unidimensional scale should have an item H below 0.3 (van Schuur 2003).

A *Crit* value is also produced using the Mokken scale analysis program for polytomous items (MSP) (Molenaar et al. 2000), which calculates a combined single value from all H coefficients for items included with evidence about the items frequency, size of model violations and significance of violations. *Crit* values of 0 are considered to indicate perfectly non-intersecting items in a scale. While *Crit* values of <40 (i.e. not 0) are thought to be the result of sampling error and values of >80 are considered to indicate violations of monotone homogeneity and double monotonicity. It is therefore acceptable to include *Crit* values between 0 and 40 (Molenaar et al. 2000; Deary et al. 2010).

A further statistic is produced by the Mokken scale analysis for polytomous items software package (Molenaar et al. 2000), the *Pmatrix*. The *Pmatrix*, shows the probability of obtaining items that are non-intersecting (Deary et al. 2010). However a value is also produced which replaces visual inspection by a 'count' of violations. When serious violations occur, it is recommended to a) consider if non-intersection (double monotonicity) is essential for the measure application under investigation or b) remove one item at a time beginning with the worst fitting and re-evaluate the scale (Molenaar et al. 2000).

Mokken analysis also produces a rho statistic which can be considered an equivalent of Cronbach's alpha for evaluation of internal consistency (Van der Lee et al. 2002; van Schuur 2003). Mokken analysis therefore also allows for the evaluation of internal consistency.

Sample size for Mokken analysis

Sample size, as in CTT methods, is important for IRT approaches. For the Rasch model, to be 95 per cent confident that no item calibration score is more than $\pm$1 logit

from its stable value, 30 participants are required for dichotomous response options. To improve on this calibration to ± ½ logit 100 participants are needed (Linacre 1994; Wright and Tennant 1996). However, evaluations of measurement items using the Rasch IRT method have generally used much larger numbers of subjects, particularly for polytomous items (Barrett and Kline 1981; Cano et al. 2008; Hobart and Cano 2009).

In Mokken analysis participant numbers of 100 to 200 and above have been recommended (Molenaar et al. 2000; Koh et al. 2006). Some authors have used participant numbers within this range (Biemans et al. 2001). However, as in Rasch analysis, studies evaluating scales using Mokken analysis have often used much larger numbers (Gillespie et al. 1987; Wismeijer et al. 2008; Deary et al. 2010). Some authors have also applied Mokken analysis in a preliminary manner with smaller numbers. Van der Putten and colleagues used it in evaluation of the Top Down Motor Milestone Test in 66 children with profound multiple disabilities and Van der Lee and colleagues applied it to evaluate the Action Research Arm Test in 63 participants (Van der Lee et al. 2002; van der Putten et al. 2005). In the van der Putten study, it was not possible to confirm double monotonicity but monotone homogeneity was identified. The number of participants was a possible limiting factor in not being able to further evaluate this measure.

3.3.3.5 Internal consistency

Internal consistency refers to the interrelatedness of a set of items (Schmitt 1996). In CTT psychometric methods, this is often attributed to homogeneity of the items. Homogeneity refers to the degree to which scale items are measuring aspects of the underlying construct or trait (Henson 2001). Measures of internal consistency should be applied to scales or sub-scales only if they are related to a single underlying construct. Internal consistency applied across multiple sub-scales would be meaningless as they are theoretically (or actually if dimensionality has already been demonstrated) not addressing a single trait. A number of methods for calculating internal consistency are widely used, such as item-total correlation (Nunnally 1978), split-half internal consistency (Nunnally and Bernstein 1994), Kuder-Richardson 20 (Kuder and Richardson 1937) and Cronbach's alpha (Cronbach 1951).

Cronbach's alpha gives the average of all split-half reliabilities of the scale. In split-half calculation, items are randomly divided into two sub-scales (for example odd and even numbered items) and the correlation between the two is then evaluated. If internal consistency is demonstrated, the two halves should be highly correlated. Limitations of this method are that there are many ways to divide a test and it does not indicate which items may not be as highly correlated. Cronbach's alpha therefore gives the average of all split halves, which is a more robust measure of internal consistency.

The utilisation of Cronbach's alpha to index the internal consistency of measures has been common practice in developing questionnaire based clinical scales (Schmitt 1996; Streiner 2003a; Barreca et al. 2005). All items addressing the same concept should at least moderately correlate with each other and have a moderate correlation with the total score (Streiner 2003a). Schmitt has discussed when coefficient alpha might be useful as well as its limitations (Schmitt 1996). He has identified four considerations or limitations for the use of Cronbach's alpha.

The first issue raised is the use of homogeneity and internal consistency, which have been used by some authors as approximately the same thing (Schmitt 1996). However, internal consistency measured by Cronbach's alpha refers to the interrelatedness of a set of items, while homogeneity refers to the unidimensionality of a set of items (already discussed in this section). Schmitt therefore indicates that in his view internal consistency is insufficient to demonstrate unidimensionality. Secondly, the appropriate level of alpha is raised as a possible area of confusion among researchers. A level of alpha above 0.70 does not necessarily indicate that a scale items are interrelated; for example, a scale of more than 20 items is likely to have a high alpha regardless of the degree of relationship between the items. Thirdly, related to the previous example and the possibility that more than one construct may be represented in the scale (Van de Winckel et al. 2006), the resulting alpha may be the product of more than one construct and still be relatively high. Fourthly, in light of the previous limitations, presenting only alpha when discussing the relationships between multiple measures is not sufficient. This view has also been echoed by authors in the rehabilitation literature (Hobart et al. 1996; Tennant and Young 1997; Davis et al. 1999; Hobart et al. 2001; Tennant 2007).

Nunnally and Bernstein proposed a criterion for Cronbach's Alpha, with an alpha of 0.70-0.90 indicating "good" internal consistency (Nunnally and Bernstein 1994). However Terwee and colleagues propose a figure between 0.70 and 0.95 and argue that higher Cronbach's alpha is often recorded in scales in which all items are clinically important and may not always indicate redundancy (Terwee et al. 2007). Another way of considering this is that items are important in this context because of the contribution they make to scale totals.

If outcome measures or sub-scales are to truly act as measures and therefore form unidimensional, interval scales, items cannot be included which do not fit the scale or dimension (Hobart et al. 2001; Hobart and Cano 2009). The measurement properties from a scaling perspective must therefore take precedence. However there is a dilemma from a clinical perspective, where specific items give useful clinical information. If these items also 'fit' the construct and scaling is demonstrated (e.g. by IRT approaches) then their retention is reasonable. However, if they do not fit the construct or fail to demonstrate fundamental measurement properties, in the context of the whole scale or sub-scale, they cannot form part of the measure (also see Section 3.2.4; page 76). An alternative may be for such items to be recorded separately as individual items, which inform the clinical picture but do not contribute to the measurement scale.

3.3.3.6 Reliability (Reproducibility)

Reliability from a CTT perspective is concerned with detecting the amount of error occurring during application of a measurement instrument (Streiner and Norman 2003). If the measurement error is relatively small (i.e. reliability is high), the results obtained can be relied upon (Streiner 2003a). Some authors (Stratford 1989; Terwee et al. 2007) have proposed that reliability should be distinguished from agreement (absolute error) and that both concepts fit under a heading of 'reproducibility'.

Terwee and colleagues (Terwee et al. 2007) propose that agreement concerns absolute error, which is how close the scores on repeated measures are, expressed in the unit of the measurement scale being evaluated. Reliability however is concerned with the degree to which different conditions can be distinguished from each other, despite measurement error (relative measurement error). Agreement is important because a

small degree of error is required for evaluative purposes in which one wants to distinguish clinically important changes from measurement error. A high degree of reliability is important for discriminative purposes to distinguish between different conditions (e.g. more or less severe disease).

Test re-test reliability is the standard CTT method of reliability evaluation and can be divided into two sub-categories of inter-rater and intra-rater reliability, when an external rater or observer applies a measure (Hicks 1999; Streiner and Norman 2003). However, if a measure is self-report then inter and intra-rater reliability do not apply and simple evaluation of test re-test reliability will be needed with completion of the measure on two occasions separated by time.

Test re-test reliability is often expressed in the form of coefficients, such as intra class correlation coefficients or Cohen's Kappa coefficient (Terwee et al. 2007). According to Terwee and colleagues for the calculation of reliability for ordinal measures, the weighted Cohen's Kappa coefficient should be used as opposed to intra class correlation coefficients (Terwee et al. 2007).

Test re-test reliability, as evaluated by Cohen's Kappa coefficient, is considered to be a more robust measure than simple percent agreement calculation, which could also be used, since it takes into account the agreement occurring by chance (Fleiss 1981). The Kappa coefficient can also be weighted to take into account agreement, which is not exact. A weighted Kappa coefficient places greater emphasis on widely different scores and less emphasis on only slight differences (Streiner and Norman 2003). Of the two possible methods used for weighting Kappa coefficients quadratic weights are recommended by Terwee and colleagues because they again give greater emphasis to large differences between scores and less emphasis to small differences (Altman 1991; Terwee et al. 2007).

There are no absolute levels for agreement as calculated by Kappa coefficient. A number of authors have provided guidelines for the interpretation of kappa coefficients; two are discussed in the context of this thesis. Landis and Koch have published guidelines for interpretation of kappa coefficient: below 0.00 = poor; 0.00-0.20 = slight; 0.21-0.40 = fair; 0.41-0.60 = moderate; 0.61-0.80 = substantial and 0.81-1.00 = almost

perfect (Landis and Koch 1977). According to Terwee and colleagues when evaluating the psychometric quality of measures, a positive rating for test re-test reliability is given when the intra class correlation coefficient or weighted Kappa is 0.70 or above in a sample size of at least 50 patients (Terwee et al. 2007).

3.3.3.7 Responsiveness and sensitivity to change

Sensitivity to change can be defined as 'the ability of an instrument to measure change in a state regardless of whether it is relevant and meaningful to the decision maker' (Liang 2000). **Responsiveness** is defined as 'the ability of an instrument to measure a meaningful or clinically important change' (Liang 2000) when change occurs and record 'no-change' when the condition is stable. Sensitivity to change and responsiveness are more recent terms used in the evaluation of measures and are considered by some authors as an extension of validity.

The term "sensitivity" can also be confusing due to its use in the area of predictive validity. It is therefore generally useful to use the term "responsiveness" for the following reasons a) the change is meaningful and b) confusion with predictive validity is avoided. If a measure is to be used to evaluate outcome following a given intervention, it must be responsive to the degree of change expected (Beaton et al. 2001). When evaluating measures for clinical application it is important and appropriate to establish that change can be recorded in a meaningful manner by the measure. The term responsiveness will be used in this thesis.

Evaluation of responsiveness will usually involve examining, either prospectively or retrospectively, the amount of change between two points in time as measured by the scale. If a prior measure is not available, retrospective scoring is possible, but is vulnerable to potential bias. Two main sources of bias of this method have been identified (Fayers and Machin 2007):

1. It is very difficult for participants to remember their past status and accurately recall it when asked to do so. The correlation in this situation between the current state and the measure is high, but is low between the measures recording of the previous state and the actual previous state (Guyatt 1987). The implication is that asking participants to do a post outcome rating of baseline is unreliable.

2. A further complication is that of response shift. Response shift means that even if accurately recalling the previous state, the participants perception of that state changes with time (Schwartz and Sprangers 1999). For example in rating pain, the participant may have rated initial pain as 8 out of 10, but having had much worse pain, now rate it as 4 out of 10.

For both of these reasons it is more reliable to score the measure at baseline and following intervention (i.e. prospectively), rather than do a retrospective score. Having acquired pre and post scores, it is then necessary to compare the scores.

There are different methods for assessing change between two measurement points on a scale. The most commonly applied methods are based on parametric assumptions within the realm of CTT and are therefore not really applicable to ordinal scales. Three methods will be discussed here due to their relevance to clinical studies (Hobart et al. 2001). Most methods are based on effect size (ES). Cohen's ES is mean change divided by the standard deviation of the baseline scores (Cohen 1988). This test is simple to apply between pre-intervention to post-intervention. A similar method is Guyatt's responsiveness, which was specifically developed for a pre-intervention, post-intervention cohort (Guyatt 1987). Another method is the standardised response mean (SRM). SRM is the ratio of the mean change (of the group) to the standard deviation of the change scores (McHorney and Tarlov 1995). An alternative non-parametric approach to descriptive presentation is the proportion of positive and negative ranks for the measure in question.

ES and SRM (and Guyatt's responsiveness) are measures of variability in results, from baseline. However, these methods have been used in evaluating many, if not most, measures developed using CTT approaches, in keeping with the assumptions of this approach. Hobart and colleagues compared effect sizes derived from total scores, with scores from the same measures derived from person locations having undergone Rasch analysis (Hobart et al. 2010). They found the responsiveness of the two measures (Barthel Index - BI and Functional Independence Measure motor scale - FIMm) was similar using total scores. But also found that FIMm had better responsiveness using a Rasch transformed scale. Nevertheless, both measures were shown to be responsive using total scores despite the limitations of using a method appropriate for interval data

on ordinal scores. Therefore, measures developed using CTT may still provide a valid record of clinical change despite their limitations.

3.3.3.8 Interpretability

The term interpretability, used by Terwee and colleagues, refers to the degree to which qualitative meaning can be assigned to quantitative scores from a measure (Terwee et al. 2007). The principal of interpretability is providing information about what change in score or point on the scale would be clinically meaningful. According to Terwee and colleagues, this requires the interpretation of scores from reference populations and thus providing normative values. In addition, a Minimal Important Change (MIC) value is also desirable, this is:

> "the smallest difference in score in the domain of interest which patients perceive as beneficial and would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management" (Jaeschke et al. 1989).

The MIC can be calculated using two methods; a criterion based method and a distribution-based method. The criterion-based method is produced by calculating the average change in the responder group. The distribution-based method, as recommended by (Norman et al. 2003), is calculated by using half the baseline standard deviation as an estimate of MIC. However, calculation of MIC using these methods requires interval scaling.

The criterion-based method is an indication of mean change in the responder group and therefore may not be a clear indication of what true MIC should be. The distribution-based method is a prediction of what meaningful change should be, and is therefore a more robust indication, although does not test this. A preliminary approach to testing may include calculation of sensitivity (proportion of actual positives correctly identified), specificity (proportion of actual negatives correctly identified), positive predictive value (proportion of those identified by the test who actually have responded to intervention) and negative predictive value (proportion of those not identified who have actually not responded to intervention) at the predicted levels of MIC, if a method

to identify individuals who respond from those who do not is possible (Lynch and Lanspa 2010; MedStats.Org 2011).

3.3.3.9 Feasibility

Outcome measures should be both practical to use in routine practice and retain psychometric properties, thus ensuring the utility of the data produced (Slade et al. 1999; Slade 2002b). Data may then be put to research purposes, but more importantly in this context, may be used to inform and direct treatment (Slade 2002).

To be feasible the measure needs to be suitable, sustained, meaningful in typical clinical settings and be applied in a specified manner for a specified purpose (Slade et al. 1999). 'Meaningful in typical clinical settings' is defined as psychometric properties being retained when the measure is used in routine practice. 'Typical clinical settings' are those where the staff are not routinely involved in research, and have no association with the measure. 'Use in a specified manner' entails identifying what is being assessed, where, when, by whom and on whom the assessment is being made. The 'specified purpose' is the identification of the use to which the information identified will be put. Feasibility is therefore important when developing a new clinical measure for application in 'normal' clinical settings, ensuring that such measures are practical and timely to apply.

3.4 Summary

A range of considerations for developing and evaluating measures have been discussed. The key issues for this thesis are:

- Theoretical models such as classical test theory, latent variable modelling and item response theory underpin our understanding of development and use of measures of latent traits. Classical test theory, latent variable modelling and item response theory have resulted in the development of specific methodologies, which enable the testing, and development of measurement scales in rehabilitation.
- Clinical imperatives can be used as a starting point for measure development. However, the use of psychometric principles in the process is desirable to ensure the resulting measure is based on sound measurement principles.

- It is essential that if a series of observations are to form a measurement scale they need to refer to the same dimension or construct. In addition, to enable fundamental measurement, observations need to conform to hierarchical interval scaling such that the summed score is a sufficient statistic.
- Rasch analysis has been developed and applies a mathematical model to scales, which if they conform to the model, confers interval level measurement.
- Initial stages of measure development may include exploration of the ordinality of measure data and evaluation of dimensionality using non-parametric item response methods such as Mokken analysis.
- For a measure to be useful in clinical practice it should be feasible, that is suitable, sustained, meaningful in typical clinical settings and be applied in a specified manner for a specified purpose. According to Slade (1999), feasibility relies on the retention of the psychometric properties of the measure when used in the clinical context.

3.5 Measure development strategy taken in this thesis

The development of the Arm Activity measure (ArmA) originates from clinical practice in that the constructs of active and passive function as concepts have been identified in focal spasticity management using BTX and PT interventions. Evaluation of the properties of ArmA needs to identify that active and passive function are separate dimensions, because without the confirmation of the sub-scales as distinct entities, measurement (rather than just describing a group of items) is not possible. The evaluation should therefore also begin to determine if items in these sub-scales can be used to effectively measure the underlying constructs or dimensions.

Possible items for inclusion in the ArmA will be identified from the literature and clinical practice sources. Initial reduction of the items will use a modified Delphi consultation followed by a wider confirmatory process. Psychometric evaluation of the ArmA will use CTT approaches and a preliminary application of Mokken analysis. Mokken analysis has been applied because of the desire to confirm unidimensionality using an IRT method in addition to a CTT method (principal components analysis). Mokken analysis allows the evaluation of raw scores for ordinality as oppose to Rasch analysis, which examines interval scaling on transformed scores. Ordinality is

particularly important in measures used in clinical practice where interpretation of raw scores may be used to evaluate patient outcome.

3.5.1 Item generation and reduction

As part of this work, identification of clinically useful and applicable items is of high importance (see Item generation; page 91). Using the literature solely as a source of items is a well-tested approach, but risks missing potentially important items. The combination, of using both the literature and analysis of agreed goals as sources, ensures that the list of possible items more fully reflects the constructs under investigation. A particular emphasis of the goal analysis was identification of passive function items, which have more limited representation in the literature, but were a particular focus of the goals analysis.

A modified Delphi consultation is then used in this thesis to select the items. Modified Delphi consultation has the advantages of anonymity to participants and reduction of the influence of socially dominant individuals (Burns et al. 2003; Finger et al. 2006). These characteristics make this consultation method well suited to obtaining individual views on measurement items from different professionals. A wider consultation with different clinicians, patients and carers is included to ensure the wide acceptance and face validity of the items selected.

3.5.2 Unidimensionality and scaling

In this thesis, CTT methods will initially be applied to development and evaluation. The broadly clinimetric approach to identification of items that are feasible combined with the wide acceptance of CTT in the literature provide strong arguments for this approach. However, the limitations of CTT approaches in terms of measurement scaling are acknowledged.

Therefore, the measurement approach used in this thesis is broadly CTT in orientation but with preliminary application of an IRT method (see Unidimensionality page 99 and Scaling page 102). Initially principal components analysis is applied to confirm the dimensionality of the two clinically conceived sub-scales. However, to ensure robust measurement principles, IRT in the form of Mokken analysis (monotone homogeneity)

was used, as a preliminary step, to identify the capacity of the sub-scales to differentiate people with ordinal scaling and to confirm dimensionality.

Monotone homogeneity only allows differentiation between people and not items; therefore reference is made to implications for double monotonicity, and ordering of items, in the future evaluation of the scale.

3.5.3 Validity

Given the importance of feasibility of the resulting measure, content validity will be addressed during measure development. Face validity will be also initially addressed but could be more formally evaluated with a specific group of patients and carers for this purpose. However, evaluation with patients and carers is integrated within pilot testing in this work and tests both content and face validity. Ecological validity will also be tested by consulting patients and carers who have undergone spasticity management intervention in the upper limb.

Construct validity will be important to ensure that items reflect the constructs of the two sub-scales and therefore also support unidimensionality. Construct validity will be evaluated by comparison with other measures, evaluating hypothesised change for convergence and divergence (see Validity page 98).

3.5.4 Internal consistency

Internal consistency in this thesis will be evaluated using Cronbach's Alpha (see Internal consistency page 105). Internal consistency is a CTT method, which could be seen as unnecessary when the application of Mokken analysis is planned. However, the application of Cronbach's Alpha supports evaluation of the resulting measures psychometric properties using CTT requirements, allows comparison with those of other measures, and for these reasons will be included.

3.5.5 Reliability (Reproducibility)

Evaluation of reliability uses a CTT perspective and is concerned with keeping measurement error to a minimum. Test re-test evaluation will be undertaken because the scale is self-rated and therefore inter-rater evaluation is not appropriate (see Reliability page 107). In ordinal scales such as the ArmA, non-parametric methods will

be required. Weighted Cohen's Kappa coefficient is the statistical test used in this thesis and takes into account agreement occurring by chance.

3.5.6 Responsiveness to change

Responsiveness will be evaluated between two time points, before and after intervention in addition to ES and SRM (see Responsiveness page 109). The primary method of determining responsiveness is by comparing the ArmA detection of functional improvement with a clinician categorisation of outcome. Amount of change is then compared between responder and non-responder groups identified by clinicians. This approach will be applied as the primary approach because ES and SRM require parametric assumptions, which are not met. The use of ES and SRM will be applied (given that these techniques require interval scaling), to allow comparison with other measures and to inform discussion of findings. Due to the acknowledged limitations of these parametric approaches, positive and negative ranks are also presented for each measure from baseline to review.

3.5.7 Interpretability

In this preliminary exploration of the ArmA, evaluation of interpretability will only be tentative, because only one reference population will be used. In addition, calculation of MIC again requires interval scaling, the evaluation of which will not be undertaken. However, initial calculation will be used to give an indication of MIC (see Interpretability page 111). This will need to be confirmed once Rasch analysis has been applied evaluating interval measurement in future work.

The MIC will be calculated using two methods; a criterion based method and a distribution-based method. Both these methods will be used to calculate MIC in the responder group for ArmA passive and active function sub-scales. These calculations of MIC will then be tested by determining sensitivity and specificity of the ArmA using those rated by clinicians to have responded to intervention or not.

3.5.8 Feasibility

Feasibility will be assessed by obtaining the views of patients and carers (see Feasibility page 112). Patients and carers will complete a questionnaire with the purpose of providing comment on the ArmA's ease of use, relevance, acceptability

and value in the clinical situation. These concepts include the time taken to complete the measure and ease with which the individual is able to undertake the completion, both of which are particularly relevant in a self-report measure.

Feasibility is difficult to evaluate in a psychometric evaluation study, where by definition practice is not normal because research is being undertaken. However, in this study, the actual impact on normal practice is considered small and the evaluation of feasibility, while having limitations, is justifiable although again preliminary.

3.5.9 Summary

The criteria used to evaluate psychometric properties will be adapted from those by Terwee and colleagues (Terwee et al. 2007) incorporating the work of the Scientific Advisory Committee (SAC) of the Medical Outcomes Trust (Medical Outcomes Trust 2002). These adapted criteria will be applied in a systematic review of upper limb function measures in this thesis (Chapter 4). The further adapted criteria presented in Table 3.1 (page 95) will be subsequently used to evaluate the ArmA measure following its development (Chapter 7). In addition, IRT methods of scale evaluation for measurement will be considered in a preliminary manner, using Mokken analysis to assess for ordinality of the scale as an initial step (See Table 3.1).

Chapter 4 Systematic review of activity measures in the upper limb

4.1 Introduction

Outcome measurement is applied in rehabilitation to determine the effectiveness of interventions. Whether in clinical practice or for research, measures need to be valid, reliable and responsive to clinically relevant change. Global measures of function in daily activities, such as the Barthel Index (Wade and Collin 1988), provide a general assessment of independence but are often unresponsive to focal interventions in the upper limb. Small changes, which may be extremely important to the patient and/or their carers are easily lost amongst the larger number of unchanging items (see Chapter 1; Section 1.3; page 40) (Ashford and Turner-Stokes 2006). Measures of active and passive function are required which capture outcome following focal interventions in the upper limb. Both active and passive function have potential for improvement following spasticity intervention with passive function improvements often more common (Bhakta et al. 2000a). Any comprehensive outcome measure for spasticity intervention needs to assess both active and passive function to fully reflect the changes seen post intervention, despite active function changes being comparatively rare.

4.2 Objectives

This Chapter will address objectives 1 and 2 of the thesis.

Objective 1. To identify standardised outcome measures of active and passive function used to assess outcome following focal intervention, in the hemiparetic upper limb, which reflect 'real-life' function.

Objective 2. To identify candidate items for inclusion in a measure of upper limb function for use following focal spasticity interventions, in the hemiparetic upper limb.

4.2.1 Search criteria:

Criteria for selection and review are summarised in Box 4.1.

Box 4.1 Summary of review criteria

Stage 1 – Selection to identify standardised outcome measures

1. Clinical relevance - applicable in the hemiparetic upper limb
2. Include items measuring:
    b. Active function

    AND/OR

    b. Passive function

Stage 2 – Real-life relevance (as opposed to under test conditions)

3. Assessed in a manner reflective of 'real-life' function

Stage 3 – Literature based evaluation of evidence for psychometric properties of measures

4. Practical to apply in everyday clinical practice
5. Valid and reliable for upper limb function evaluation
6. Responsive to change occurring as a result of intervention

4.3 Method

The review was undertaken in three stages.

- In stage 1, a 'pool' of possible measures was identified from a broad-based systematic search.
- In stage 2, these were reduced to those reflective of 'real-life' performance.
- In stage, 3 measures were evaluated based on the published evidence for psychometric properties.

The Quality of Reporting of Meta-analyses (QUOROM) provides guidance on the most appropriate methods of presenting systematic review data and these principles were used in the presentation of data and results (Moher et al. 1999).

The review was confined to neurological rehabilitation for three reasons:

1) Measures identified outside of the neurological rehabilitation literature have limited applicability to this population due to the impairment pattern presented with hemiplegia (Wade et al. 1983).
2) Many of the items included in measures designed for general rehabilitation populations, are often too difficult for the minority of patients with hemiplegia who re-gain active function with the upper limb. The result of item difficulty is that improvements are not demonstrated even when clinical change is taking place (floor effect).
3) A joint initiative of the American Academy of Orthopaedic Surgeons, the Council of Musculoskeletal Specialty Societies and the Institute for Work and Health (Toronto, Canada) conducted a literature review of currently available measures of function and disability in the upper limb. The review identified no current measures which were *"self administered, able to assess functional status for upper extremity musculoskeletal conditions and were feasible for use in research or routine clinical settings"* (Hudak et al. 1996; Davis et al. 1999) p1. As a result the Disability of the Arm, Shoulder and Hand measure (DASH) was developed. The literature review conducted;

though not a systematic review was detailed and wide-ranging in identifying musculoskeletal self-report measures for upper limb function available. Other reviews have been undertaken since, such as that by Bot and colleagues (Bot et al. 2004), which focused on the impact of shoulder impairment on function, however this identified the DASH as the most useful measure of function in the upper limb for musculoskeletal presentations. The DASH remains the most relevant 'real-life' measure of active function in a non-neurological, rehabilitation population, but was not designed for individuals with neurological impairment.

4.3.1 Stage 1 Measure selection

Criteria for selection and review are summarised in Box 4.1. In Stage 1, standardised outcome measures were identified for further consideration if they were:

a) Applicable in the hemiparetic upper limb.
b) Included items measuring active and/or passive function.

Data sources

The Cochrane Database of Systematic Reviews was investigated to identify databases and other data sources to be searched for this review. The following data sources were then searched:

1. Medline search based on the strategy outlined by Dickersin (Dickersin et al. 1994). (1996 to 7th May 2008).
2. CINAHL (1982 to 7th May 2008).
3. BIDS Science Citation Index (1991 to 7th May 2008).
4. Embase (1974 to 7th May 2008).
5. Relevant trials were identified in the Specialised register of Stroke trials. (1993 to 7th May 2008).
6. National Health Service National Research Register, MRC Clinical Trials directory, Database of Abstracts of Reviews of Effects (DARE), Google, ProFusion and SIGLE (medical/rehabilitation grey literature), (until 7th May 2008).

7. The Cochrane Database of Systematic Reviews (1993 to 7th May 2008).
8. Reference lists from papers identified.
9. Conference proceedings, books and book chapters.
10. Communication with lead authors of published studies and other researchers.

Search strategy

The search strategy for the clinical group (stroke or brain injury) was;

1. (hemiplegia or hemiplegic or hemiparesis or hemiparetic)
2. AND (arm or upper limb or hand or shoulder)
3. AND (stroke or post stroke or CVA)
4. OR (brain haemorrhage or haemorrhage or haematoma or hematoma)
5. OR (brain injury)
6. OR (brain tumour or tumor)
7. OR (brain infection or encephalitis or abscess)

The further search strategy for the outcome measurement sub-group was;

8. AND (Outcome measurement [MESH] OR Outcome assessment)
9. AND (function* OR activity).

Recommended by the Cochrane Collaboration (Dickersin et al. 1994).

The search strategy for data sources 3, 4, 5 and 6 (see Data Sources) was modified to that given above due to the more limited search capacity available in those search tools. Single search term searches were undertaken and then combined to allow full searching in those data sources. Search terms were altered when required to comply with 'key' terms used in other databases as appropriate.

The search strategy for data source 7 (see Data Sources), also involved a modified and simplified search strategy from that used in 1 and 2.

Reference lists and textbooks were searched by hand (Data Sources 8 and 9). Textbooks were either identified in the wider search (as indicated above) or through searching the catalogues of a number of medical libraries or through discussion with expert clinicians in rehabilitation medicine or physiotherapy. Conference proceedings were identified from the search strategies applied in 1 and 2 searching medical libraries and discussion with expert clinicians or researchers as well as searching key textbooks.

Where appropriate, as indicated by the literature or through discussion with other clinicians or researchers, published authors and researchers (n=6) were contacted about the scales that they had used or developed (see Appendix 1) (Data Source 10). Contact was usually by e-mail or telephone and was followed up a maximum of twice if an initial response was not obtained.

The title was reviewed to identify potentially relevant studies, and then the abstract was reviewed where available. When the abstract indicated relevance or no abstract was available, the full text paper was retrieved and a final decision made about inclusion. Initial selection was undertaken by the author and was then evaluated by a second reviewer; any areas of disagreement were discussed. Publications selected were restricted to those in the English language.

4.3.2 Stage 2: Real-life relevance

Measures were excluded if they did not use a method of assessment reflective of 'real-life' function to measure day-to-day performance. In Stage 2, selected measures were considered to have **'real-life' relevance** if they assessed day-to-day performance in the person's normal environment, as opposed to performance when observed under test conditions (such as a standardised test in a clinic setting).

4.3.3 Stage 3: Evaluation

The names of measures identified in stage 2 were used as terms for a further search of the electronic databases to obtain original and any subsequent publications concerning their development and psychometric evaluation. Medline, CINAHL and the reference lists of identified publications containing relevant outcome measures were then searched to identify further literature on the development of these outcome measures and their

psychometric properties. Authors of outcome measures were contacted for further details when required.

Based on this published literature, the psychometric properties of each measure were evaluated against the following review criteria:

- **Practicality for use in everyday practice:** time to complete, burden, readability
- **Validity and reliability:** content validity, internal consistency, construct validity, floor and ceiling effect, test-retest reliability, agreement
- **Responsiveness to change:** demonstration of change following focal upper limb intervention, interpretability and minimal important change (MIC)

Descriptive information was tabulated for each of the selected measures including; the items in the measure, the methods of administration and the method of scoring applied. The quality criteria developed by Terwee et al (Terwee et al. 2007) and used by Bot et al (Bot et al. 2004) for a "clinimetric evaluation of shoulder disability questionnaires" from those produced by the Scientific Advisory Committee (SAC) of the Medical Outcomes Trust (Medical Outcomes Trust 2002) were used to evaluate the quality of the instrument properties. The full criteria used with minor alterations to those produced by Bot and colleagues are given in Appendix 2 and the application in this instance is described in assessment methods below. Two reviewers independently evaluated each measure using these criteria. Findings were then compared and any discrepancies resolved through discussion. The option was available for a third reviewer to resolve any areas of disagreement following comparison, but was not used.

Data sources

The following data sources were then searched:

1. Reference lists from papers identified above
2. Medline (1996 to 7th May 2008).
3. CINAHL (1982 to 7th May 2008).
4. Communication with lead authors of published studies and other researchers.

Search strategy

The search strategy

1. Measure name AND
2. Psychometric evaluation
3. OR testing
4. OR validity
5. OR reliability
6. OR application
7. OR clinical application

Procedure used to evaluate each measure

Psychometric properties were rated for each measure, using a scale of Adequate (+), Doubtful (±), Poor (-), or Unknown (?).

Administrative burden

Administrative burden was assessed using the same scoring method, modified as follows: Easy (+), when dichotomous items were simply summed; Moderate (±), when an ordinal or visual analogue scale was used to quantify individual items then summed, and Difficult (-) when a summary score was applied in combination with a formula. Timing for completion of the measure was also rated as positive for measures completed within 10 minutes.

Validity

The instruments were evaluated for content and construct validity on the scale used for all psychometric properties. A positive rating for content validity was given when there was evidence that either patients, carers or other experts had been consulted regarding the initial selection of items (e.g. through focus groups or surveys) or had provided evaluation or feedback as part of the development. A positive rating for construct validity was given if there was evidence that the measure was based on hypothetical constructs, which had been tested and supported during its evaluation.

Internal consistency

A positive rating for internal consistency was given if the factor structure of the measure had been tested through factor analysis, or where ratings for Cronbach's alpha were between 0.70 and 0.95 for each dimension or subscale.

Floor and ceiling effects

Floor and ceiling effects were considered present if more than 15% of respondents achieved the highest or lowest possible score respectively. Floor effects were also considered present if the measure only assessed bilateral or complex tasks (e.g. thread a needle or sharpen a pencil).

Reliability

Test-retest reliability was rated as positive if repeat testing of the same condition had yielded comparable results e.g. an intraclass correlation coefficient (ICC) of greater than 0.70 for total scores. In item-by-item analyses, agreement was also rated as positive if it had been evaluated and shown to be satisfactory, using accepted statistical methods such as the Kappa coefficient or standard error of measurement.

Responsiveness

Responsiveness was rated as positive if the measure had demonstrated significant change in response to intervention, in the context of an appropriate study design (see full criteria Appendix 2).

Interpretability

Interpretability (see Glossary) is the degree to which qualitative meaning can be assigned to quantitative scores (Nunnally and Bernstein 1994). Positive ratings were given if at least two types of information were given to aid in understanding of the scores. Information considered included, means and standard deviations of the score totals before and after treatment, information in relation to other clinical variables, which might be expected to change, or information on the minimum change in score that might be clinically meaningful using the MIC.

4.4 Results

4.4.1 Stage 1: Measure selection

The search yielded 1,144 studies, including primary reports, abstracts and conference proceedings.

- Eighty-four papers were identified following initial review of the abstracts as including measures of functional outcome following focal upper limb intervention, yielding a total of 20 outcome measures after stage 1. A summary of the stages of review, according to QUOROM, is given in Figure 4.1.

The initial evaluation of these 20 measures is shown in Table 4.1. This includes identification of psychometric evaluation in the published literature, but does not indicate the quality of the evaluation carried out.

Figure 4.1 QUOROM Measure selection flow diagram

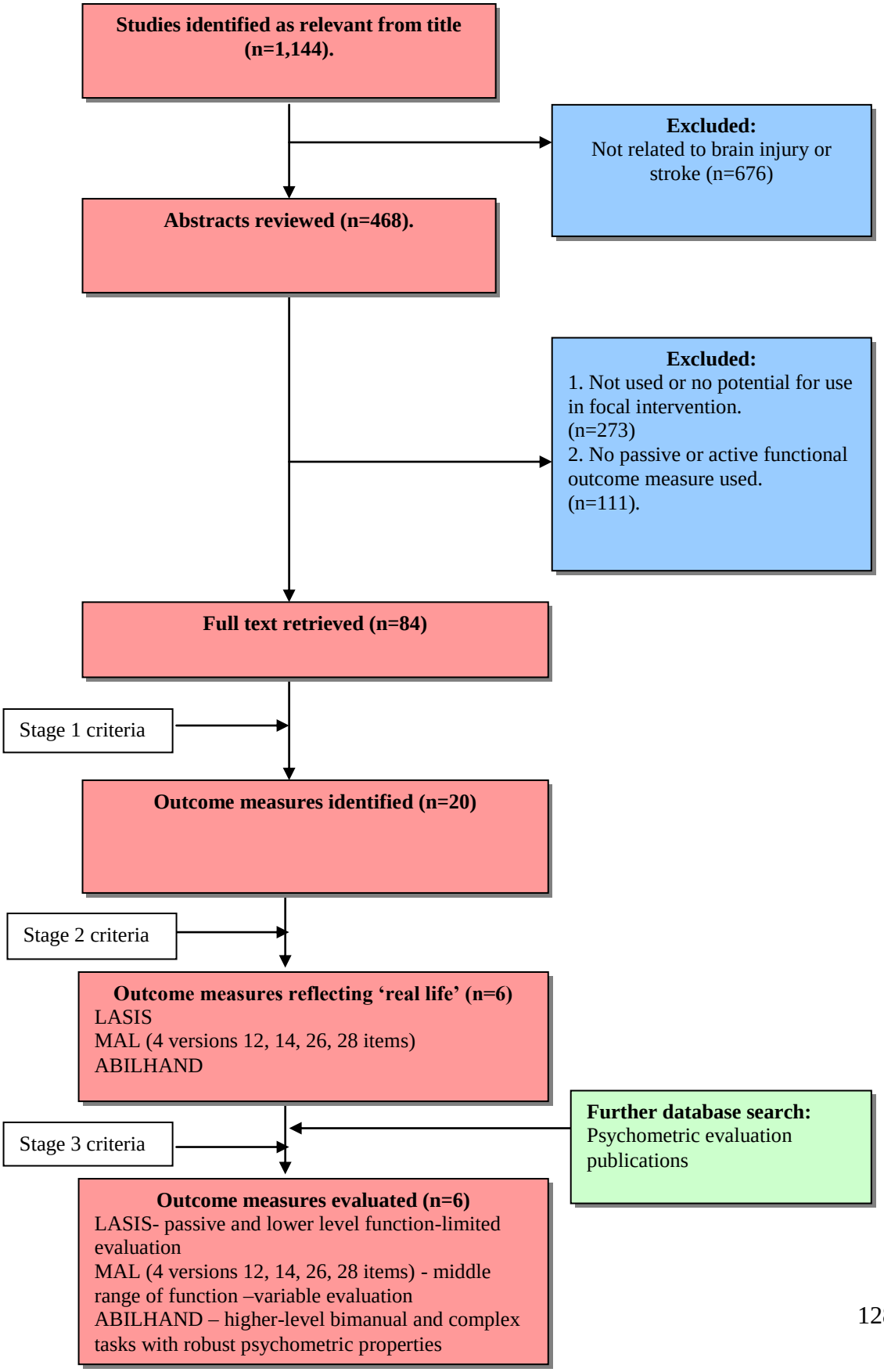


Table 4.1 Identified outcome measures

No	Outcome measures	Reflective of real-life	Apply hemi-paretic upper limb	Active function elements	Passive function elements	Evidence of formal psychometric testing in neurological rehabilitation
1	Leeds Adult Spasticity Impact Scale (LASIS) (Bhakta et al. 1996; Bhakta et al. 2000a)	√	√	√	√	
2	Disability Assessment Scale (DAS) (Brashear et al. 2002a; Brashear et al. 2002b)		√		√	√
3	Motor Activity Log (MAL-14) (van Kuijk et al. 2002; Page et al. 2003; van der Lee et al. 2004; Yelnik 2004; Dettmers et al. 2005; Ring and Rosenthal 2005)	√	√	√		√
4	Motor Activity Log (MAL-12) (Popovic et al. 2003)	√	√	√		√
5	Motor Activity Log (MAL-26) item (van der Lee et al. 2004)	√	√	√		√
6	Motor Activity Log (MAL-28) (Uswatte et al. 2006)	√	√	√		√
7	ABILHAND (Penta et al. 1998; Penta et al. 2001)	√	√	√		√
8	Wolf Motor Function Test (Wolf et al. 2001; Page et al. 2003; Cusick et al. 2005; Dettmers et al. 2005)		√	√		√
9	Box and Block Test (Desrosiers et al. 1994; Alon et al. 2003; Lannin and Herbert 2003; Higgins et al. 2005)		√	√		√
10	Action Research Arm Test (Wade 1992a; Boiteau et al. 1997; Broeks et al. 1999; Parry et al. 1999; Page and Levine 2003)		√	√		√
11	Frenchay Arm Test (Wade 1992b; Alon et al. 1998; Lagalla et al. 2000)		√	√		√

No	Outcome measures	Reflective of real-life	Apply hemi-paretic upper limb	Active function elements	Passive function elements	Evidence of formal psychometric testing in neurological rehabilitation
12	Rivermead Motor Assessment Arm Scale (Wade 1992b; Parry et al. 1999)		√	√		√
13	Nine Hole Peg Test (Wade 1992b; Lindberg et al. 2004; Ring and Rosenthal 2005; Rodgers 2008)		√	√		√
14	Upper Extremity Function Test (Popovic et al. 2003)		√	√		√
15	Motor Club Assessment (Wade 1992b)		√	√		√
16	Jebson Hand Function Test (Wade 1992b; Jones et al. 1996; Richardson et al. 1997; Lannin and Herbert 2003)		√	√		√
17	Fugl-Meyer Upper Limb Test (Berglund and Fugl-Meyer 1986; Wade 1992b)		√	√		√
18	Purdue Peg Board Test (Wade 1992b; Desrosiers et al. 1995; Hurvitz et al. 2003)		√			√
19	Arm Motor Ability Test (Kopp et al. 1997)		√	√		√
20	Chedoke Arm and Hand Activity Inventory (Barreca et al. 2005)		√	√		√

Key: √ Indicates that the attribute is present, without any assessment of quality.

4.4.2 Stage 2: Real-life relevance

Six measures were identified, which met both Stage 1 and Stage 2 criteria (i.e. were relevant to real life functional performance). These were the LASIS, the ABILHAND and the MAL, which had four different versions MAL-14, MAL-26, MAL-28 and MAL-12 (see Table 4.1). All four versions of the MAL had undergone separate psychometric evaluation so were included. The scaling methods, number of items and methods of administration for these measures are shown in Table 4.2.

Table 4.2 Selected measures of function

Outcome measure	Method and procedure of scoring	Context for development
Leeds Adult Spasticity Impact Scale (LASIS) (Bhakta et al. 1996; Bhakta et al. 2000a)	**Items:** 12 **Scoring:** Patients and carers, over the past 7 days. Items rated between 0-4. Scores summed and divided by the number of questions answered. **Administration:** Semi-structured interview.	Spasticity (BTX) intervention
Motor Activity Log (MAL-14) (van Kuijk et al. 2002; Page and Levine 2003; van der Lee et al. 2004; Yelnik 2004; Dettmers et al. 2005; Ring and Rosenthal 2005; Uswatte et al. 2005)	**Items:** 14. **Scoring:** by patients, over the past 7 days. Scores range from 0 (never use the affected limb for this activity) to 5 (always use the affected arm for this activity). Subjects are rated on the amount they use their paretic arm ("amount scale") and on the quality of their movement during the functional activities ("how well scale"). **Administration:** Semi-structured interview.	Constraint Induced Movement Therapy (CIMT)
26 Item MAL (MAL-26) (van der Lee et al. 2004)	**Items:** 14 original items, 11 additional items and 1 optional item chosen by the patient. **Scoring:** by patients, over the past 7 days as per the MAL. **Administration:** Semi-structured interview.	CIMT
28 Item MAL (MAL-28) (Uswatte et al. 2006)	**Items:** 28 **Scoring:** by patients, over the past 7 days or past 3 days. **Administration:** Semi-structured interview.	CIMT
12 Item MAL (MAL-12) (Popovic et al. 2003)	**Items:** 12 **Scoring:** by patients, over the past 7 days as per the MAL. **Administration:** Semi-structured interview.	CIMT
ABILHAND (Penta et al. 1998; Penta et al. 2001)	**Items:** 23 **Scoring:** Patients asked to estimate their ease or difficulty of performing each task (without help) only on tasks they have performed. **Administration:** Semi-structured interview.	Chronic Stroke rehabilitation

4.4.3 Stage 3: Evaluation

The detailed evaluation of the properties of the selected measures is presented in Table 4.3, using the upper limb specific quality criteria developed by Bot and colleagues (Bot et al. 2004).

Table 4.3 Quality assessment of selected measures based on analysis of the published studies

Measure	Time	Admin. Burden	Content Validity	Internal Consistency	Construct Validity	Floor / Ceiling Effect	Reliability	Agreement	Responsiveness	Interpretability
LASIS	+	±	?	?	?	?	?	?	?	?
MAL-14	-	+	?	+	±	±	-	-	-	±
MAL-26	-	+	?	+	±	±	-	?	-	?
MAL-28	-	+	?	+	±	±	±	±	?	?
MAL-12	±	+	?	?	?	?	?	?	?	?
ABILHAND	-	+	+	+	+	-	+	+	+	+

Method or result was rated as: + Adequate; ± Doubtful; - Poor; ? No data available.

Quality assessment based on the criteria by Bot et al (2004).

Table 4.4 provides a summary of the methodological quality of the specific psychometric evaluation studies for each of the selected measures.

Table 4.4 Summary of the methodological quality of the psychometric studies of selected measures using the COSMIN checklist

Outcome measure	Methods Applied	Measurement properties addressed in the paper	Comments on methodological quality
Leeds Adult Spasticity Impact Scale (LASIS) (Bhakta et al. 1996; Bhakta et al. 2000a)	No published formal evaluation	Content Validation	Using an open interview with patients and carers to create and item bank. • Details of the methods for this not published. Second study used the measure in an intervention study.
Motor Activity Log (MAL-14) (Uswatte et al. 2005)	CTT	Internal consistency Construct validity Responsiveness	Examination in 41 stroke participants. • Internal consistency using Cronbach's alpha was demonstrated for both sub-scales • Construct validity evaluation was inconclusive. • Limited methods to correlate change for responsiveness (Pearson Correlation) Limited evaluation of psychometric properties.
26 Item MAL (MAL-26) (van der Lee et al. 2004)	CTT	Internal consistency Test-retest reliability (Bland and Altman method) Cross-sectional construct validity Longitudinal construct validity (responsiveness)	Clinimetric evaluation in 56 chronic stroke patients. • Internal consistency demonstrated in both sub-scales • Test-retest reliability measurements 2 weeks apart but used Bland and Altman in MAL data which is ordinal. • Correlation of the two sub-scales with each other and with the total score but not with another measure. • Longitudinal construct validity evaluated by responsiveness ratio (mean change after intervention) non significant responsiveness demonstrated.

Outcome measure	Methods Applied	Measurement properties addressed in the paper	Comments on methodological quality
28 Item MAL (MAL-28) (Uswatte et al. 2006)	CTT	Item analysis Internal consistency Construct validity	Comparison of MAL-28 scores with accelerometer findings in 222 stroke patients. • Item analysis using item total correlations resulted in removal of two items. • Internal consistency demonstrated. • Construct validity involved comparison with accelerometer recordings and a subjective hand Function Scale but had limitations with accelerometer comparison and application.
12 Item MAL (MAL-12) (Popovic et al. 2003)	CTT (clinimetric)	Content Validation	Clinimetric reduction of items from 14 to 12, no further evaluation of psychometric properties.
ABILHAND (Penta et al. 1998; Penta et al. 2001)	IRT (1 Parameter model) Rasch Analysis	Hierarchical scaling properties Content validation Construct validation Unidimensionality Reliability (error measure variance)	Initial evaluation in 18 rheumatoid arthritis patients (Penta, 1998). Preliminary evaluation using Rasch analysis. Subsequent evaluation on 103 chronic stroke patients (Penta, 2001). More expansive robust evaluation using Rasch analysis.

In table 4.4 reference has been made to the COSMIN checklist for assessing the methodological quality of studies of the measurement properties of health status questionnaires (Mokkink et al. 2010). However, these criteria were not available at the time of the original review and therefore the Bot (2004) upper limb specific criteria were used in Table 4.3.

Table 4.5 shows the item content of each of the six identified measures can be broadly placed in a hierarchy of increasing difficulty.

Table 4.5 Items included in each measure

Functional Items	**LASIS**	**MAL-14**	**MAL-26**	**MAL-28**	**MAL-12**	**ABIL-HAND**
Passive Function Items						
Cleaning the palm affected hand	**1**					
Cutting fingernails affected hand	**2**		**25***			**4***
Cleaning the affected elbow	**3**					
Cleaning the affected armpit	**4**					
Cleaning the unaffected elbow	**5**					
Putting arm through coat sleeve	**6**	**1***	**1***			
Difficulty putting on a glove	**7**					
Difficulty rolling over in bed	**8**					
Doing physiotherapy exercises to arm	**9**					
Active Function Items						
Difficulty balancing standing	**10**					
Difficulty balancing walking	**11**					
Hold object steady, use other hand (jar[a])	**12**					**10[a]**
Steady myself while standing		**2**	**2**			
Carry an object from place to place		**3**	**3**	**23**	**12**	
Pick up fork or spoon, use for eating		**4**	**4**	**24**	**10**	
Comb hair		**5**	**5**	**25**		
Pick up cup by handle		**6**	**6**	**26**	**11**	
Hand craft/card playing		**7**	**7**			
Hold a book for reading		**8**	**8**			
Use towel to dry face or other body part		**9**	**9**			
Pick up a glass		**10**	**10**	**20**	**5**	
Pick up toothbrush and brush teeth		**11**	**11**	**21**	**6**	
Shaving / make-up		**12**	**12**			
Use a key to open a door		**13**	**13**	**22**	**7**	
Letter writing/typing		**14**	**14**		**8**	
Pour coffee / tea			**15**			
Peel fruit or potatoes			**16**			**3**
Dial number on the phone			**17**			
Open / close a window			**18**			
Open an envelope			**19**			
Take money out of a wallet or purse			**20**			
Undo buttons on clothing			**21**			
Buttons on clothing (shirt[a], trousers[b])			**22**	**27[a]**		**13[a] 17[b]**
Undo a zip			**23**			

Functional Items	**LASIS**	**MAL-14**	**MAL-26**	**MAL-28**	**MAL-12**	**ABIL-HAND**
Do up a zip (jacket [a], trousers [b])			**24**			**11[a] 21[b]**
Other optional activity			**26**			
Turn on a light with a light switch				**1**		
Open a drawer				**2**		
Remove item of clothing from drawer				**3**		
Pick up phone				**4**	**1**	
Wipe kitchen counter				**5**		
Get out of car				**6**		
Open refrigerator				**7**		
Open a door by turning a door knob				**8**	**2**	
Use a TV remote control Wash your hands				**9** **10**		
Turn water on/off with faucet (tap)				**11**	**4**	
Dry your hands				**12**		
Put on your socks				**13**		
Take off your socks				**14**		
Put on your shoes				**15**		
Take off your shoes				**16**		
Get up from chair with arm rests				**17**		
Pull chair away from table before sitting				**18**		
Pull chair toward table after sitting				**19**		
Eat half a sandwich or finger food				**28**	**3**	
Use removable computer storage					**9**	
Hammer a nail						**1**
Thread a needle						**2**
Wrap gifts						**5**
File nails						**6**
Cut meat						**7**
Peel onions						**8**
Shell hazel nuts						**9**
Open pack of chips (crisps)						**12**
Sharpen pencil						**14**
Spread butter						**15**
Fasten ‘snap’ (press stud)						**16**
Cap of a bottle						**18**
Open mail (post)						**19**
Squeeze toothpaste						**20**
Unwrap chocolate						**22**
Wash hands						**23**

Key:

Items in the table are given the number at which they appear in order in the measure.

Items in LASIS included under passive function all asked respondents ‘how difficult’ a task was to undertake related to care of the limb by the patient him or herself or a carer.

* Items in the passive function section included in MAL-14, MAL-26 or ABILHAND could be done either passively or with more active involvement by the individual, with the focus being on active involvement in these measures.

[a] and [b] refer to specific objects used for the functional items in a measure.

At the lowest level of the hierarchy in Table 4.5, the LASIS, includes mainly passive function items. In the middle of the hierarchy, the MAL contains items of active

function, increasing in the following order: MAL-14, MAL-26, MAL-28, and MAL-12. At the upper level of the hierarchy, the ABILHAND contains complex items often requiring bilateral hand use, and for ABILHAND, the order of difficulty has been confirmed by Rasch analysis (Penta et al. 1998; Penta et al. 2001).

Administrative burden and time for completion

The administrative burden was adequate for all measures apart from the LASIS. The LASIS requires the calculation of the measure total, however this calculation in practice is not complex. The calculation of the LASIS involved totalling the item scores and calculating the mean. There is no lower limit to the number of items, which need to be answered for a valid global score, so the score across different individuals may not be comparable.

All measures are completed by the clinician using a structured interview with the patient and carer and require allocation of clinician time. The LASIS, three versions of the MAL (excluding MAL-12), and the ABILHAND were thought to involve a time for completion of greater than 10 minutes, although specific data on this were not available.

Validity

Construct validity was adequately addressed in three versions of the MAL (14, 26 and 28) and ABILHAND. Information on floor and ceiling effects was difficult to identify or not formally addressed in the majority of measures. However, given that the measures have a hierarchical relationship in their item content, it may be expected that the LASIS would have ceiling effects in a higher function group, and similarly the MAL and ABILHAND would have floor effects for detecting changes in lower level and passive function tasks.

Reliability

Internal consistency was demonstrated in four measures; three versions of the MAL (14, 26 and 28) and ABILHAND. Test-retest reliability evaluation was documented in four measures the ABILHAND, the MAL-14, MAL-26 and the MAL-28 (van der Lee et al. 2004; Uswatte et al. 2006). Adequate methods have been used in the ABILHAND, but were less convincingly applied in the MAL-14, MAL-26 and MAL-28.

Responsiveness

Responsiveness was demonstrated in the ABILHAND and was also assessed in the MAL-14 and 26. However, the change in the MAL-14 and 26 did not correspond to change identified by other measures, and responsiveness was therefore rated as inadequate in this evaluation (van der Lee et al. 2004; Uswatte et al. 2005). Responsiveness in both measures was evaluated in post stroke hemiplegic patients who had good return of arm movement and related function.

Interpretability

Interpretation of specific scores with respect to qualitative meaning had only been evaluated in the MAL-14 and ABILHAND. The ABILHAND had been evaluated using Rasch analysis and demonstrated a clear gradation of increasing ability of different items within the scale (Penta et al. 1998; Penta et al. 2001). It was therefore given a positive rating. The MAL, however, did not show an adequate relationship between overall scores or achievement of individual items and qualitative meaning. The MIC was not clear and it was therefore given a doubtful rating overall.

4.5 Discussion

This systematic review identified six measures (including four versions of MAL), which had been used in the published literature to evaluate function reflective of real-life or actual performance. The six measures appeared to fall broadly into a hierarchy of increasing difficulty. The LASIS evaluates passive function and low-level active function, such as using the affected hand to hold and stabilise objects. The MAL and ABILHAND were more comprehensive (and consequently complex) measures for active function, evaluating a wide range of activities, including unilateral and bimanual function.

In terms of their psychometric properties the LASIS and MAL-12 have received limited evaluation and met only one of the Stage 3 review criteria each. The MAL-14, 26 and MAL-28 have been more extensively validated, but although they each met two criteria, their performance was doubtful on the remainder. Only the ABILHAND has been

thoroughly evaluated and was shown to meet 9 of the 11 criteria, although it failed on time for completion and floor effects in a more dependent group.

The implication of these findings for clinicians is that there are several measures available and the choice of measure will depend on the patient's current level of function and the anticipated goals for treatment.

- The LASIS is likely to be useful for individuals who have little or no active movement or function, but nevertheless have care and maintenance issues related to the hand and upper limb.
- The MAL-14 contains more unilateral and simple items, which may be useful for detecting change in individuals who have some (but limited) arm function.
- The MAL-26 also includes the 14 items but adds a further 12 -including some tasks (such as peeling potatoes or taking money out of a purse) which require two hands.
- The MAL-28 includes seven items from the MAL-14/26, but adds a further 21 functional tasks, some of which challenge reach and strength (such as putting on shoes and socks or pulling a chair towards a table after sitting), while the MAL-12 represents a short version that spans the entire range of MAL items
- The ABILHAND has six items in common with these scales, but adds a further sixteen, all of which are more complex bilateral tasks. It is therefore likely to be useful for patients functioning at a higher level (see Table 4.2 for details of the measures and Table 4.5 for the included items).

4.5.1 Limitations:

The systematic review has limitations in three main areas: identification of items, missing studies and evaluation of psychometric properties.

Identification of measures

The starting point for the review was the scientific literature, and it is possible that measures have been missed that are used in clinical practice, but have not been applied in research. However, as the objective was to identify measures for which there is some evidence of psychometric evaluation, it is considered appropriate to base the review in the research literature.

The possibility of missing studies

The secondary search for literature regarding psychometric evaluation included identification of references from the original publications and a search of the cited literature based on the name(s) of the measures. The LASIS has not been referred to by this name in the literature, but as disability and carer burden scales in BTX intervention studies by the Bhakta and colleagues (Bhakta et al. 1996; Bhakta et al. 2000a). The current version of the LASIS was obtained by directly contacting the author after the initial search. As highlighted in the case of the LASIS, it is possible that the narrower secondary search may have missed some of the grey literature. However, it was anticipated that these other publications would generally be of lower quality, and would not add significantly to the body of evidence that was found.

Evaluation of psychometric properties

The use of formal evaluation criteria supported a detailed assessment of the published psychometric properties for the respective measures. The criteria published by Terwee et al (Terwee et al. 2007) were based on an earlier review by the same group (Bot et al. 2004). The criteria were not developed for the context of hemiplegia, although the earlier review was a systematic review of shoulder disability questionnaires for application following musculoskeletal injury. The current review did not identify any of the same measures evaluated by Bot, due to the different patient populations considered. For example, Bot identified the DASH (Hudak et al. 1996) questionnaire which best met their search criteria and had undergone the most extensive psychometric evaluation. The DASH is a self-report measure of everyday active function. Self-reporting may have advantages in reducing the clinical time required to administer the measure. However, it is designed to assess higher-level function and, like the ABILHAND, is likely to show floor effects in a neurologically impaired population. At the other end of the scale, none of Bot's measures contained any passive function items. Passive function applies particularly in the context of neurological damage, but could also have relevance in very severe musculoskeletal conditions, such as deforming arthritis. This emphasises the wide range of functional activities of the upper limb.

The results of the review can also be compared to those of an Occupational Therapy (OT) orientated review by Rowland and Gustafsson (Rowland and Gustafsson 2008).

Their review, published after the systematic review in this thesis, considered upper limb measures of activity as defined by the ICF for specific application in occupational therapy practice. The OT orientated review involved searches of Medline, the Cumulative Index to Nursing and Allied Health Literature (CINAHL) and the Cochrane library. This was a less extensive search than that conducted for this thesis, which included searches of other databases based on the strategy outlined by Dickersin and colleagues (Dickersin et al. 1994). Rowland and Gustafsson's review was not systematic in nature, which in part explains the smaller number of sources searched. The review criteria were similar in focussing on measures of activity, though differed in not specifying measures that reflect everyday performance. They used quality criteria proposed by Law and Baum (Law and Baum 2001) rather than those by Terwee and colleagues (Terwee et al. 2007) used in this review.

The criteria by Law and Baum address the focus of the measure, clinical utility, scale construction, standardization, reliability and validity. Explicit definitions for each criterion were not given with the degree of detail provided by Terwee, although the evidence for each measure relating to individual criteria was included. Some of the measures identified were the same as those identified in this systematic review; Wolf Motor Function Test; Arm Motor Ability Test; Action Research Arm Test; Motor Activity Log (14 and 28), Upper Limb – Motor Assessment Scale; ABILHAND and Chedoke Arm and Hand Inventory. The LASIS was not identified because it is referred to by different titles in the literature, it was specifically developed for evaluation of spasticity and does not have published psychometric evaluation.

4.6 Implications and conclusions

This systematic review of measures identified a selection of validated tools available for the evaluation of 'real-life' active function in the hemiparetic upper limb. None provide a comprehensive assessment of both active and passive function. Depending on difficulty of the goals for treatment, clinicians could select from the six measures presented in this review but would need to be aware of the limitations in psychometric evaluation for some of these measures as discussed.

The ABILHAND appears to be a robust measure for higher levels of function, and the range of different versions of the MAL allow for assessment of abilities in the middle range. However, there is a specific gap in relation to measures that assess passive and lower level function. Moreover, all of the measures identified in this review are administered by structured interview, which has implications for clinician time if used in routine clinical practice. The development of self-completed questionnaires has the potential to improve the practicality of application, although some patients with neurological disability may find this difficult, especially if they have significant cognitive or communicative problems. Further exploration and development of measures to address these issues is required.

An alternative to the development of a single new measure, addressing these deficiencies is the development of an item bank of the type referred to by authors such as Tennant (2007) using Rasch analysis techniques for development. Such an approach would use the six measures identified in this review and the preliminary hierarchy presented in Table 4.5 as a basis for development. There are potential strengths in having a range of items which fully capture the extent of active function. In addition the measure would also have potential application in both spasticity intervention and in other focal upper limb interventions, in which, active function improvement is much more likely.

However a weakness of item banking for this work is evident with limited exploration of passive function items having been undertaken in the development of the current measures. The majority of change following spasticity intervention is expected in passive function rather than active function and therefore this issue is particularly important. This will be addressed further in Chapter 5. In addition, for active function items, development of an item bank would make the assumption, that all relevant items are included in the existing measures without evaluating this. This will also be evaluated further in Chapter 5.

In a minority of cases (possibly those with more recent injury) improvement in active function may still also occur following focal spasticity intervention. It may therefore be important to capture improvements in both active and passive function. This presents a

difficulty in selecting items from the item bank for clinical application, where active function items may not be applied because change is expected only in passive function. If active function change does occur no baseline measure of this sub-scale will have been recorded.

One way of ensuring that passive function is appropriately measured and that changes in active function are detected if they occur, is initially to develop a measure which has two sub-scales addressing both active and passive function. Such a measure could, in due course, form the lower end of an item bank (or two item banks for active and passive function) but would also ensure initial clinical utility in spasticity management practice.

In summary, there is a need for a self-report measure of passive and active function in the upper limb. Chapter 5 identifies further items for inclusion in such a measure.

Chapter 5 Identification of patient selected items

5.1 Introduction

The systematic review yielded a range of possible measurement items, which primarily addressed active function rather than passive function. This chapter complements the systematic review by involving patients and carers at an early stage of measure development, focusing on identifying passive function items as well as confirming the importance of items already identified from the literature. The analysis in this chapter involves the evaluation of actual clinical goals set by patients and carers in conjunction with the clinical team. The purpose of this work is to ensure that items of particular relevance to patients and carers, particularly those for passive function, are reflected in the development of the new measure.

Setting goals with patients and carers has been identified as a core activity in rehabilitation practice (Playford et al. 2009; Wade 2009). The application of goal setting in directing intervention is established practice for most neurological rehabilitation professionals and within most service settings. Although currently much discussion is taking place as to the exact nature and purpose of goals in rehabilitation practice (Hart and Evans 2006; Hurn et al. 2006; Latham and Locke 2007), the use of some form of goal setting is widely accepted (Wade 2009). A rehabilitation goal is defined as the aim or target of a specific intervention or programme of interventions (Wade 1992a). In practice, goals are often set at a number of different levels from completion of small tasks such as sitting in a wheelchair for 30 minutes, to whole rehabilitation programmes with the aim of being independent and self-caring at home. The emphasis in rehabilitation practice is on making goals functional and meaningful to the patient. Goals are therefore usually directed at achieving change in either the activity or participation levels of the ICF (Wade 1992b; WHO 2002). Intervention may be aimed at changing impairments to body structures, but the primary focus is likely to be functional improvement. Goal attainment scaling (GAS) is one method used to set goals, which has the advantage of enabling the quantification of the outcome related to the goal in a systematic manner (Kiresuk and Sherman 1968).

The GAS technique is suitable for health problems, which warrant an individualised approach to outcome evaluation (Stolee et al. 1992; Rockwood et al. 1997; Stolee et al. 1999; Zaza et al. 1999). GAS has been used in a number of rehabilitation studies (Stolee et al. 1992; Rockwood et al. 1997; Stolee et al. 1999) and has also been used following spasticity intervention (Ashford and Turner-Stokes 2006; McCrory et al. 2009; Turner-Stokes et al. 2010). In addition, application of GAS has been recommended for evaluation of individual patient goals in the recent guidelines for management of spasticity with botulinum toxin (Royal College of Physicians et al. 2009). A full description of the GAS method is given in Appendix 3.

When evaluating function in the arm, much of the clinical focus has been on the hand and distal upper limb. The hand and distal upper limb perform the tasks essential to arm function, such as manipulating and using objects and tools. Many of these tasks are assessed in existing outcome measures (Penta et al. 1998; van der Lee et al. 2004). The proximal upper limb and shoulder also have an important role in positioning the upper limb to perform many manipulative tasks. Shoulder and elbow movement is also important for passive function tasks such as cleaning the armpit or elbow crease and these functions are not as frequently addressed in existing measures. This was felt to be an important omission.

The work presented in this chapter was a secondary analysis of a cohort study previously published (Ashford and Turner-Stokes 2008) (see Appendix 20) and followed the systematic literature review (Ashford et al. 2008). The review yielded a range of possible items to be included in a new measure, but revealed a preponderance of active function measurement items, with far fewer addressing passive function, and indicating the need for further research to determine a core set of items for inclusion in a new measure. A key secondary aim was to identify proximal as well as distal upper limb items.

This secondary analysis therefore has three aims:

1. Identification of new passive function items by patients and carers.
2. Confirmation of items identified in the systematic review.

3. Ensuring that items relevant to the proximal upper limb are considered for inclusion.

5.2 Objective

Objective 2 was addressed in Chapter 4 and will also be addressed in this chapter.

Objective 2. To identify candidate items for inclusion in a measure of upper limb function for use following focal rehabilitation interventions, in the hemiparetic upper limb.

5.3 Method

The analysis used goals for spasticity intervention involving combined BTX and focal PT intervention (physical interventions specific to the upper limb, such as splinting).

5.3.1 Design

The study used a prospective observational cohort design. Ethical permission for the study was obtained from the Harrow Research Ethics Committee – EC2773 (see Appendix 10).

5.3.2 Setting

Physical therapy (PT) interventions and BTX were provided according to a previously described ICP (see Appendix 11) for spasticity management as standard clinical practice through a tertiary spasticity service (Ashford and Turner-Stokes 2006). Injection of BTX used a flexible protocol according to the different muscles involved, together with concurrent PT interventions, consisting of splinting, serial casting, exercise programmes, functional electrical stimulation, arm supports and patient and carer education on arm positioning and stretching.

5.3.3 Selection of participants

Participants were identified from a tertiary service resulting in a sample with more severe disability than in typical stroke populations and to include proximal upper limb problems as a particular focus.

Inclusion criteria:

- Hemiplegic upper limb impairment affecting function.
- Impairment resulting in an increase in muscle tone and alteration to strength and control.
- Undergoing treatment for spasticity management in the shoulder girdle or proximal upper limb requiring BTX intervention and physical therapy.
- Age between 18 and 85 years.

Exclusion criteria:

- Patient declines to participate or family and/or treating team declines on their behalf.
- Unable to complete goal setting process and no carer (professional or family) available to undertake completion. Examples of situations that may lead to exclusion are indicated below:
    - Does not speak English
    - Unable to communicate responses (where feasible, communication using adapted methods was employed, including support from a speech and language therapist to avoid exclusion).
    - Cognitively unable to understand process.

The sample comprised a consecutive cohort of patients referred for proximal upper limb spasticity management with BTX and physical therapy intervention. All presented with spasticity following either brain injury or stroke and had been referred to the service by their physiotherapist, general practitioner or medical consultant. Carers of patients were also consulted about goals for intervention when they were involved in providing care.

5.3.4 Procedure

Patients were provided with information before agreeing to participate in the study (see Appendix 4). Consent to participate in the study was obtained and a written record kept (see Appendix 5). All functional goals were determined through discussion between patient and/or their carers and the clinical team (state registered health professionals). Goals were set using GAS (Turner-Stokes 2009b) using the standard method and were entered into the ICP proforma to ensure appropriate recording and capture. Goal

achievement was also recorded on the clinical ICP document (see Appendix 11 for the current version of the ICP). A second clinician reviewed the goal setting procedure, to ensure that goals had been recorded for each patient and were appropriate to the clinical presentation.

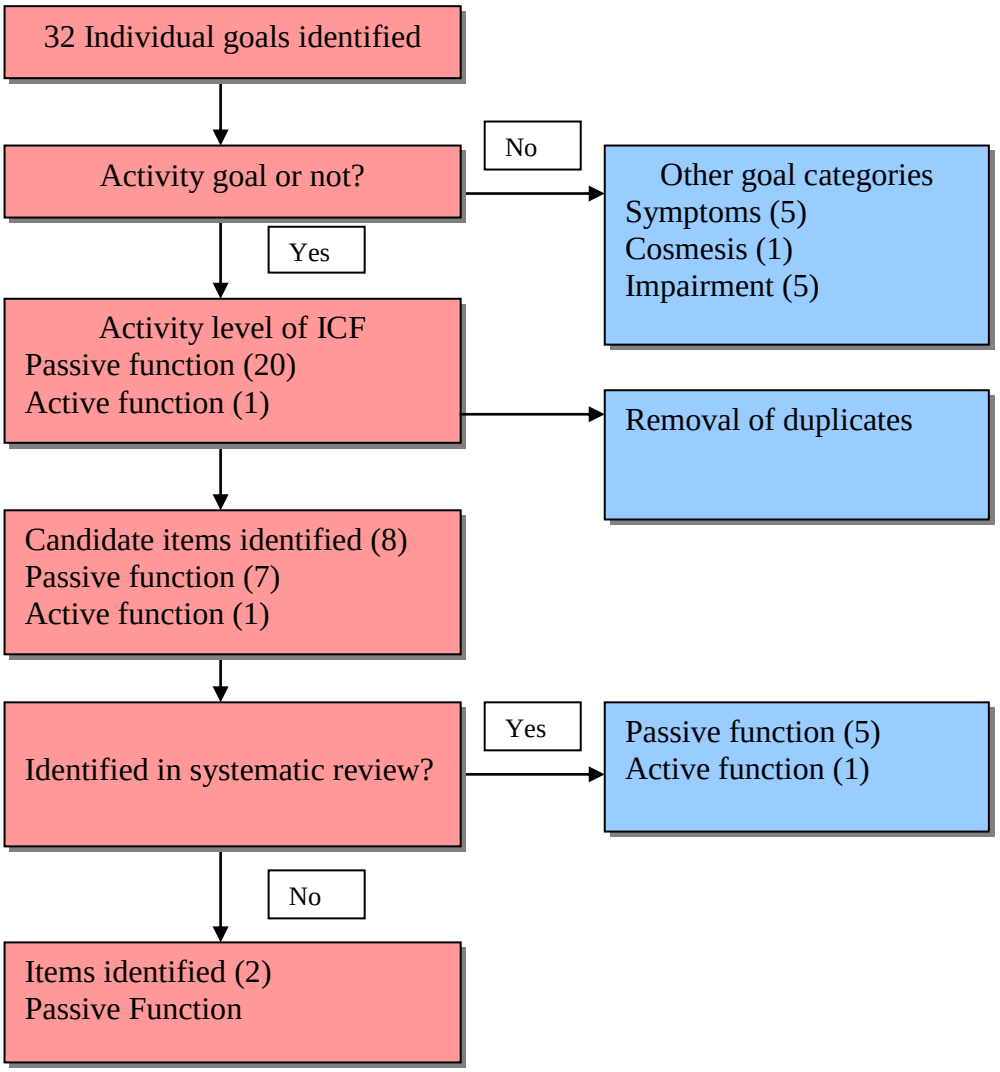
Goals were reviewed to identify categories. Goal category identification, followed by assignment of goals to a category, was undertaken by the author and then reviewed by a colleague. Active or passive function goals were retained and others were excluded. Each goal was then considered as a potential item for the ArmA. Goals with slightly different wording but representing the same issue were collapsed into a single item. The resulting list of items was compared with those identified in the systematic review. Items not already identified from the literature were included in the development of the ArmA.

5.4 Results from goal setting analysis

A total of 78 patients were treated for upper limb spasticity, of which 16 had treatment for shoulder girdle or proximal involvement and were included in the secondary analysis. All consecutive referrals undergoing BTX intervention for proximal upper limb spasticity gave consent to be included as participants. The mean age of the 16 patients was 54.5 years (SD 15.7), 9 were male and 7 female, and diagnostic groups were brain injury due to trauma (n=1) and stroke (n=15). The mean age is at the younger range of ages in the general stroke population. The median baseline Barthel Index was 10.5 (Inter-quartile range 5 to 18), from a total possible score of 20, which indicated a moderate level of disability in this group. Mean time since neurological injury was 15.7 (SD 20.9) months.

Intervention goals were reviewed and then allocated to one of five categories; passive function, active function, symptoms (e.g. pain), cosmesis and impairment (see Figure 5.1).

Figure 5.1 Process of item selection



Thirty-two goals were set in all for the 16 patients, two goals per patient (see Table 5.1).

Table 5.1 Goals set by each participant (n=16).

Patient	Goals	Goal category				
		Passive Function	Active Function	Symptoms	Cosmesis	Impairment
1	Reduce shoulder pain			√		
	Allow independent selective shoulder movement to enable reach for cup		√			
2	Enable positioning of right upper limb with reduced pain	√		√		
	Enable washing and dressing of upper limb with reduced pain.	√		√		
3	Reduction in difficulty dressing due to reduced shoulder pain	√		√		
	Reduction in difficulty washing under arm due to reduced arm pain	√		√		
4	Improved ease of dressing arm	√				
	Improved ease of washing arm	√				
5	Reduction of shoulder pain			√		
	Improved ease of positioning arm	√				
6	Enable tolerance of wrist and hand splint	√				
	Enable positioning of upper limb	√				
7	Improve positioning of upper limb for improved postural control in sitting	√				
	Improved cosmesis of upper limb				√	
8	Prevent further loss of range of movement					√
	Enable splint application	√				
9	Enable palmar hygiene	√				
	Improve ease of dressing upper limb	√				
10	Reduction in shoulder pain			√		
	Increase ease of axillary hygiene	√				
11	Prevent deterioration in range of movement					√
	Prevent elbow hygiene difficulties	√				
12	Ease of splint application	√				
	Prevent loss of range at the elbow					√
13	Reduce associated reaction from moderate to mild			√		
	Reduce wrist pain by two points on visual analogue scale			√		
14	Maintain current range of movement					√
	Maintain current ease of care	√				
15	Prevent further loss of range of movement at the elbow					√
	Maintain elbow crease skin hygiene	√				
16	Increase ease of washing left shoulder	√				
	Increase ease of dressing left shoulder	√				
Totals		**20**	**1**	**9**	**1**	**5**

The option was available to set more than two goals (3 to 4 goals are recommended for spasticity intervention in recent guidelines (Royal College of Physicians et al. 2009)). However, in practice only two were set per patient in the 16 patients included in this study based on the GAS discussions between patients, carers and the clinical team. The setting of no more than two goals may have occurred due to the relative novelty of the GAS procedure or time restriction in the clinic environment. However, goals set for patients 2 and 3 corresponded to two goal categories. When reduced to a list of candidate items, eight items were identified relating to the activity level of the ICF, comprising seven passive function and one active function. The items were then compared with the systematic review. Two passive function items had not been identified by the systematic review and were then included in the item reduction process to develop the ArmA (see Table 5.2). The two items were splint application and placement of the upper limb on a support (wheelchair tray).

Table 5.2 Passive function items identified by participants (n=16)

Identified items	**Number of times this goal was set**	**Identified by systematic review**
1. Washing upper limb including shoulder.	**4**	√
2. Dressing upper limb including shoulder.	**4**	√
3. Axillary hygiene.	**2**	√
4. Elbow crease hygiene.	**2**	√
5. Palmar hygiene.	**1**	√
6. *Splint application.	**3**	
7. *Enable placement of upper limb on a support (wheelchair tray).	**2**	

Key: * Items included in item reduction for the ArmA

5.5 Discussion

Goals were allocated to categories, goals in irrelevant categories were deleted, and duplicate goals were conflated to produce a list of candidate items. These items were included for item reduction in the development of the ArmA and support content validity of the subsequent measure. The use of this methodology has enabled inclusion of patient and carer selected items in the measure development process.

The five passive function and single active function items identified in both the systematic literature review and goal-setting review have been supported as relevant for use in measurement. These items have face validity evidenced by their selection by patients and carers. The construct validity for the subsequent development of the ArmA is also supported by the use of items identified in pre-existing measures and in this review of goal setting.

Of the two additional items not identified in the literature, only one 'Enable placement of upper limb on a support (wheelchair tray)' involved the proximal upper limb and shoulder. The other item of 'upper limb splinting', focused on the wrist and hand. Taking all eight items identified, five involve the proximal upper limb or shoulder and were passive function. Based on this finding, most relevant items are present in currently available measures. However, the process ensures that passive function items relevant to the proximal upper limb have been considered in development as intended.

5.5.1 Strengths and limitations

This analysis had three primary functions; 1) identification of new passive function items by patients and carers for inclusion in the ArmA, 2) confirmation of items identified in the systematic review, which enhances ecological validity and construct validity and 3) ensuring that items relevant to the proximal upper limb were considered for inclusion. These three aims were met.

Passive function items

In a sense, it may be surprising that more passive function items were not identified, in part demonstrating that patients have similar goals for focal spasticity intervention. An alternative possible explanation maybe that using goals, as a source of items could tend to produce more components that are homogeneous. In general, items measuring the same underlying construct are desirable in a sub-scale. Theoretically, this could lead to a measure which has limited ability to discriminate between patients of different ability (Gillespie et al. 1987). However, it is more likely that the key passive function items were identified and that others had already been identified by the systematic review.

Active function items

Active function goals are less frequently set in spasticity management intervention and active function improvements are only rarely seen, but do sometimes occur. Therefore, specific purposive sampling would have been unlikely to identify new items in spasticity management. Evaluating goals set for other interventions, such as 'constraint induced movement therapy, could be undertaken, but are less likely to identify items particularly relevant to spasticity intervention. In addition, the specific purpose of this work was to identify passive function items to redress the imbalance in the literature, which primarily focuses on active function items. Alternative methods could have been employed (see Item generation; pages 91), such as focus groups or interviews, but again may not have added to items identified in the systematic review.

Sample

A further limitation is that data were available on only 16 participants, all of whom had very specific problems related to proximal spasticity. While this is a strength, in potentially identifying items relevant to the shoulder and proximal upper limb, it may restrict identification of other items because the participant's clinical problems are so similar. In addition, the sample used was relatively small and all taken from one centre. However, the fact that only two new passive function items were identified indicates a degree of 'saturation' in the items already identified by the systematic review, despite being fewer than for active function.

Comparing this sample (a relatively young group with proximal upper limb spasticity) to the general stroke population, the participants were at the younger end of the age range of patients with stroke (mean age 54.5). Patients with cognitive difficulties which prevented them participating in the goal setting process in any way were excluded, as well as those who were unable to communicate responses due to significant communication impairment. However, considerable effort was expended in supporting communication and cognitive impairment to prevent exclusion. Support for such patients included communication using adapted methods such as visual aids and support from a speech and language therapist. The criticism could be made that the ArmA may not be suitable for older adults or those with communication difficulties. However, the approach taken has attempted to ensure that communication and cognitive difficulties, while certainly a barrier, are controlled for as far as is possible. The impairments and disabilities of the patients are also likely to be the more significant factors than pure chronological age in ensuring that the measure is appropriate and applicable.

User involvement

Patient and carer involvement in research and measure development has been emphasised in the rehabilitation literature and this review has facilitated this involvement in the ArmA development (Playford 2008; Giordano et al. 2009). However, a limitation of using goal analysis, is that patients and carers are setting goals of relevance to them, which may not allow them to reflect on the broader application of measurement items to other people. As emphasised by INVOLVE (the Department of Health body for public and patient involvement in health and research) (INVOLVE 2009), patients and carers have unique insights into research. In this regard, the study may have been strengthened by further consultation in the form of qualitative methods to generate items with patients and carers, in addition to reviewing personally identified goals.

Comparison with other work

Turner-Stokes and colleagues reviewed the goals for treatment in a randomised controlled trial of spasticity management using BTX (Turner-Stokes et al. 2010). The review indicated that GAS was an effective method of evaluating change following BTX intervention. However, goals were often not achieved, possibly due to over-

optimistic active function goals being set. More importantly, when goals were categorised according to the ICF, 28% (46) were set within domains related to impairment of body function, with the remaining 72% (119) goals related to the domains of activity or participation (WHO 2002; Takahashi 2004). The majority of goals however related to activity. Activity goals were identified in the following categories:

- Upper limb activities – such as lifting and carrying or holding objects still (generally not achieved)
- Mobility – e.g. maintaining balance or improving gait
- Self-care tasks such as hygiene, dressing or feeding
- Domestic and community tasks such as housework or recreational activities

It should be noted that the patient group used in the above study were more able than that used for the review of goals in this Chapter, and a difference was expected because of this. However, all the passive function items identified in the current study correspond to self-care tasks in that by Turner-Stokes and colleagues (2010), with no additional items identified by them. This supports the inclusive nature of items identified by the work in this chapter.

5.6 Conclusions

In summary, the aim, of undertaking an investigation of goal setting for spasticity intervention, to identify passive function items, was achieved addressing a possible gap in the literature. Secondary aims of considering proximal spasticity and comparison with systematic review findings were also achieved. The patient group, used in this goal analysis was, undergoing spasticity intervention and therefore had few active function goals as expected.

The approach used in this chapter, has provided a complementary and innovative approach to that of the systematic review, thus enhancing validity as well as identifying additional items for inclusion in the development of the ArmA. The analysis was particularly useful for identification and confirmation of passive function items. Chapter 6 will describe how the candidate items identified in Chapters 4 and 5 were used to develop the ArmA.

Chapter 6 Development of the ArmA measure

6.1 Introduction

The need for a new self-report measure of active and passive function for application in the hemiparetic upper limb has been identified. In this Chapter, the development of the measure is reported. Items are derived from measures identified in the systematic review (Chapter 4) and the patient identified items (Chapter 5).

6.2 Objectives

This Chapter describes the method for addressing objectives 3 and 4.

Objective 3. To develop a self-report measure to assess both 'active' and 'passive' function in the hemiparetic upper limb following focal rehabilitation interventions.

Objective 4. To confirm face and content validity by investigating item relevance for professionals (content), patients and carers (face and content).

6.3 Methods of ArmA development

The ArmA was developed using a modified Delphi Consultation, followed by wider consultation with other clinicians, patients and carers and piloting. A sub-scale was developed for both active and passive function. The theoretical concepts of active and passive function are based on the discussion in Chapter 1.

A Delphi approach uses an iterative consultation to measure opinion from identified experts (Burns et al. 2003). The pre-existing items from the systematic review and the patient-identified items, were used as a starting point for the process, rather than initial generation of items using the Delphi technique.

Modified Delphi consultation was selected because it provides anonymity to participants and reduces personality based influences such as the impact of socially

dominant individuals on the consensus process (Burns et al. 2003; Finger et al. 2006). Finger and colleagues consider the Delphi method to have four key characteristics:

- anonymity for those participating;
- iteration of concepts;
- statistical group response based on frequency of selections (in this instance item selection); and
- informed input from expert participants (Finger et al. 2006).

The literature provides no definitive recommendation on panel size, which have ranged greatly in different studies between 10 and 1685 (Reid 1988) and in the rehabilitation literature from 15 (Raine 2006) to 263 (Finger et al. 2006). Raine suggests that good results can be obtained with between 10 and 15 panel participants where the group is homogenous, and that smaller groups such as this are also more likely to retain group members (Raine 2006).

6.3.1 Ethics and Research & Development (R&D) approval

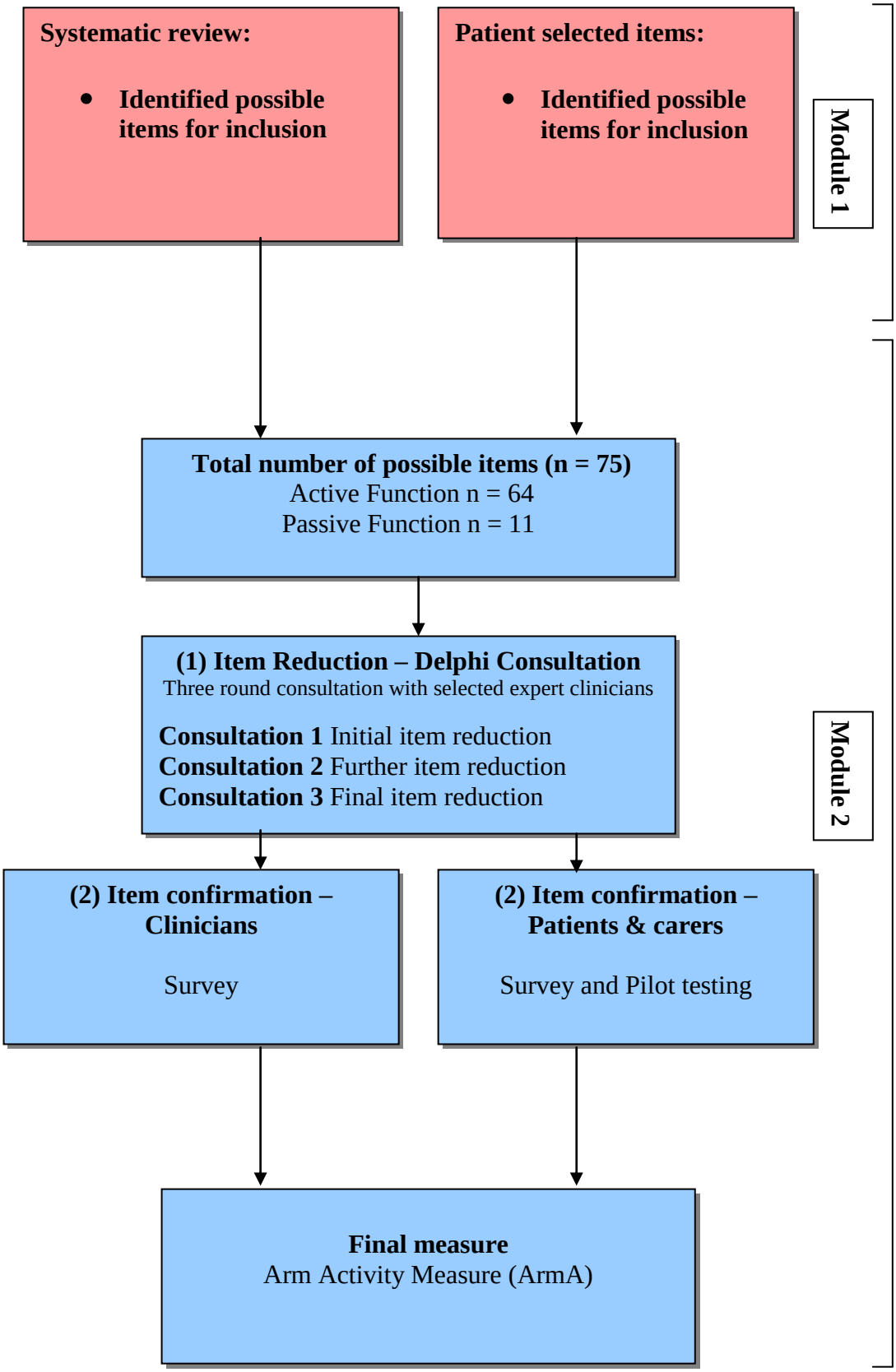
Ethical approval for all three modules of the research programme was received, from Central Office for Research Ethics Committees (COREC), now the National Research Ethics Service (NRES) by the Research Ethics Committee (Oxfordshire REC A) at John Radcliffe Hospital, Oxford (COREC number 05/Q1604/110) shown in Appendix 10. Site-specific assessment for all three modules of the study was carried out and approved by Northwest London Hospitals NHS Trust, (Northwick Park Hospital). The Research and Development department also gave approval on behalf of Northwest London Hospitals NHS Trust and agreed to be sponsor for the project.

The psychometric evaluation and cohort study required the inclusion of an additional site added to the COREC approval, the Alderbourne Rehabilitation Unit, Hillingdon Hospital, London. A Principal Investigator was identified at this site and a Site-Specific Assessment undertaken and Research and Development approval was obtained. See Appendix 10 for ethical approval documentation.

6.3.2 Summary of development

The process of the ArmA development is summarised in Figure 6.1.

Figure 6.1 Summary of ArmA development



6.3.3 Participants

Stage 1 (reduction of items) involved modified Delphi consultation with a purposive sample of experienced clinicians. Stage 2 (confirmation of items) involved wider consultation for confirmation of items selected with a different group of clinicians in addition to pilot testing with patients and carers.

Stage 1. Reduction of items using modified Delphi consultation

Participating clinicians (n=10) worked in two regional rehabilitation units, two district rehabilitation services and a community rehabilitation team within the London Region. The sample therefore included a spectrum of clinicians with experience of assessment in the upper limb from a range of clinical services. The panel of clinicians included physiotherapists (n=4), occupational therapists (n=4) and rehabilitation medicine physicians (n=2). All the therapists were either clinical specialist or senior level and the rehabilitation medicine physicians included were both consultant level.

Stage 2. Confirmation of items by wider clinician involvement and pilot testing with patients and carers

Clinicians

The group consisted of specialist physiotherapists, occupational therapists and rehabilitation nurses none of whom had been involved in earlier stages of development or evaluation. The invited physiotherapists were all identified from the UK Physiotherapy Adult Spasticity Forum, with the consultation document sent to the whole membership (n=58). All physiotherapists in the forum were involved in spasticity management services managing the upper limb following BTX intervention. Occupational therapists were identified through initial contact with the physiotherapists and worked with them in specialist neurological rehabilitation services, with involvement in spasticity management. Rehabilitation nurses were identified from rehabilitation services in North West London NHS Trust and worked with patients with upper limb activity limitation following stroke and brain injury.

Pilot testing: Patients and Carers

Patients (n=13) and carers (n=13) were identified from those receiving inpatient, outpatient or outreach spasticity management input from North West London, Hertfordshire and Bedfordshire through the Regional Rehabilitation Service.

6.3.4 Procedure

The following section describes the procedure for each stage of development comprising; initial item reduction using modified Delphi consultation, wider consultation with clinicians and pilot testing with patients and carers. The process utilized consultation documents electronically or paper based dependent on the participant's preference.

Stage 1. Item reduction using Modified Delphi Consultation

In Delphi Consultation, consensus is deemed to have been identified when the votes from respondents fall within a pre-defined range. For example, Raine (2006) used an 80% response level for acceptance of an item, where 80% of respondents agree with the item or change to the item. A range of acceptance levels can be found in the literature between 55 and 100 percent (Deane et al. 2003; Powell 2003). The COSMIN study group (see page 85) used 67% agreement between experts in the group as a cut-off point for agreement (Mokkink et al. 2010). This cut-off point was arbitrarily chosen based on evaluation of the levels of agreement seen between members of the group.

In this study, the level was set at 60% consensus for exclusion or inclusion of items. The level of agreement was set before data collection and analysis at a level thought to allow agreement between group members and to enable the reduction of items. Item reduction was a key focus of this work as well as the confirmation of the content (patients, carers and professionals) and face (patients and carers) validity of items. Reduction of items from the original 75 was needed to ensure that the measure developed was feasible in normal clinical practice. Reduction of items (or responses to Delphi questions) is an appropriate use of the technique and often involves the re-presentation of findings from earlier rounds of the consultation process (Strauss and Ziegler 1975; Burns et al. 2003; Deane et al. 2003; Powell 2003). This principal was used in the Delphi consultation in this work and the 60% cut off point for inclusion of items was used based on levels recommended by Deane et al (2003).

Three rounds of Delphi consultation were planned to enable the feedback of comments to the group. Item prioritisation by group participants, in light of the feedback, was then possible to enable decisions on inclusion or exclusion of remaining items. The approach to initial item reduction in this thesis was to use a clinical prioritisation approach within the Delphi framework to prioritise items based on clinical opinion. This approach was shown to be effective in the development of the Quick DASH, resulting in a shorter, clinically feasible measure with items prioritised by clinicians thought to have greater face validity (Beaton et al. 2005).

Delphi Consultation Round 1

The list of possible items, from the systematic review and patient identified items (see Tables 4.5; page 137 and 5.2; page 153) was presented to the purposively selected sample of expert clinicians. See Appendix 6 for the 'Item selection grid' including the list of items. The list was distributed by post or electronic mail. Respondents were asked to identify: (a) items which were important to include in a measure of active and passive arm function from the list; (b) items from the list, which should be excluded along with the reason for exclusion; (c) any items that were not on the list which were of particular importance and explain why they should be considered for inclusion. Once the comments had been returned, participants were, where necessary, contacted to clarify any points and ensure no issues had been missed. The initial list of items was revised in light of these findings to produce a short list for round 2.

Delphi Consultation Round 2

The revised list was then returned for further comment and verification, consisting of the original list and the revised short list. Respondents were asked to comment again on the items repeating the previous process. They were also asked to consider which of the remaining items were unlikely to apply in a majority of cases and would therefore not be as relevant to include. Once the comments had been returned, participants were, where necessary, contacted to clarify any points and ensure no issues had been missed. A further revision to the short list was produced in light of round 2 findings.

Delphi Consultation Round 3

The results from round two of consultation were sent out again to the same group, consisting of the original list (round 1) and the further revised short list from round 2. The respondents were asked to confirm the selection of items, with the full list of possible items available for reference. Once the comments had been returned, participants were, where necessary, contacted to clarify any points and ensure no issues had been missed.

Following third round consultation, a draft measure was constructed using the items identified by the group. Based on the findings of the systematic review the method of scoring items was adopted from the six measures selected for psychometric evaluation (see Table 4.5; page 137). The method comprised completion based on activity over the preceding 7 days and was scaled on a five point ordinal scale. This method of scaling responses was adopted as the method for the **draft ArmA**.

Stage 2. Item confirmation - wider review by clinicians

Consultation was then undertaken through e-mail or post with physiotherapists, occupational therapists and nurses who formed the clinicians for the item confirmation group. Item confirmation was undertaken because a main focus of the Delphi consultation had been reduction of items and further confirmation of content validity in a larger group of clinicians would strengthen and reconfirm the findings. In addition the wider consultation allowed for the inclusion of nurses who were a professional group not included in the Delphi consultation and enabled this possible limitation to be addressed.

The multistage approach to development of the ArmA was in part modelled on that used in development of the DASH. The initial consultation for DASH was undertaken with clinical experts and methodologists by reviewing items from relevant measures in use (Hudak et al. 1996). This was then followed by item reduction using the same group to remove items which were not relevant for various reasons defined by the group. Further item reduction was then undertaken by field testing the items with a small number of patients. The development of the Quick-DASH then involved further item reduction using three methods: the concept-retention method, the equidiscriminative item-total correlation and IRT – Rasch analysis (Beaton et al. 2005). Multiple stages of item reduction and confirmation were shown to be of value in the DASH work and contributed to the resulting face, content validity and utility of the measure developed. A similar multi-stage approach was therefore used in the development of the ArmA, with an aim to strengthen face, content validity and utility.

The consultation document consisted of the draft ArmA and the original list of items from the systematic review (see Table 4.5, page 137) and patient-identified items (see Table 5.2; page 153). The patient and carer version of the item confirmation questionnaire is shown in Appendix 7. Respondents were asked to identify: (a) items not included in the draft ArmA from the original list which should be included; (b) items included in the draft ArmA, which should be excluded along with the reason for exclusion; (c) any items that were not included in the draft ArmA or the original list which should be included and explain why. Respondents were also asked to comment on the way in which items were scaled.

Clinicians who had not returned the consultation document within two weeks were contacted again and a new consultation document sent where required. If they had not returned the consultation document following a further two weeks, they were contacted a third time, after which time follow-up was discontinued.

Patients and Carers - item confirmation

Patients were provided with information before agreeing to participate in the study (see Appendix 8). Consent to participate in the study was obtained and a written record kept (see Appendix 9).

Patients and carers were also asked to comment on draft ArmA items by responding to the same consultation document as the item confirmation group of clinicians. The patient and carer version of the item confirmation questionnaire is shown in Appendix 7. As the measure is designed for self-report or structured interview, patients and carers were also asked to complete the ArmA, and give their views on; (a) its relevance to them; (b) ease of completion (c) presentation style. Consultation documents were distributed to patients and carers, either face-to-face, returned by post or over the telephone. If patients or carers had not returned the document after two weeks they were then contacted and an additional document sent if needed. If they had not returned the document after a further two weeks, they were contacted a third time, after which time follow-up was discontinued.

The responses were then compared with the modified Delphi consultation results. If new items were presented these were considered provided they were identified by more than one respondent, either clinician, patient or carer. The researcher reviewed all comments and made decisions on changes to ArmA based on (1) issues raised by multiple respondents, or (2) issues corresponding to findings from the systematic review. Decisions about items were then discussed with a colleague for concordance before changes were made. This process resulted in version two of the ArmA for psychometric evaluation.

ArmA item mapping onto the ICF

To further demonstrate the content validity of the ArmA (version-2) it was mapped onto the ICF. The aim was to confirm the domains addressed in the ArmA by comparison with sub-categories of the ICF. The mapping was undertaken using the online ICF illustrated library available from the International University of Health and Welfare (Takahashi 2004).

Cieza and colleagues identified that with the development of the ICF its concurrent use and comparison with health outcome measures would be necessary and particularly relevant in rehabilitation to enable identification of relevant measures for practice (Cieza et al. 2002). To enable the classification of health status questionnaires such as

the ArmA using the ICF, Cieza and colleagues produced rules or criteria for the classification process (Cieza et al. 2002). These linking rules were used in the classification of the ArmA. The author and a colleague undertook the classification separately and blind to each other's initial classification. The results were then compared and any differences discussed. Agreement was achieved for classification of all ArmA items, but the option was available to ask a third researcher to adjudicate should it have proved impossible to reach agreement.

6.4 Results

The results for reduction of items using modified Delphi technique at stage one, the confirmation and pilot testing of ArmA at Stage two and the ICF classification are presented below.

6.4.1 Stage 1 - Reduction of items

Delphi Consultation Round 1

All 10 clinicians initially approached returned the round one consultation document. Following round one 48 active function items were excluded and 4 passive function items. Consensus for exclusion was between 60 and 100% (6-10 clinicians). Table 6.1 shows the initial short list of items following round one. The table also shows the measures from which the items originate or identifies that they were patient selected (Chapter 5) and the broad anatomical region of the arm addressed by each item.

Table 6.1 Initial short list of passive and active function items (round 1), mapped back onto the other measures.

Functional Items	Patient Identified	LASIS	MAL-14	MAL-26	MAL-28	MAL-12	ABIL-HAND	Proximal, Distal, Whole arm
Splint application	*							Whole arm
Positioning the arm comfortably	*							Whole arm
Putting on a glove		*						Distal
Cutting fingernails		*						Distal
Cleaning the armpit		*						Proximal
Cleaning the palm		*		*				Distal
Putting arm through coat sleeve or dressing the arm		*	*	*				Whole arm
Eat with a knife and fork			*	*	*	*		Whole arm
Pick up a glass, bottle, or can			*	*	*	*	*	Whole arm
Brush teeth			*	*	*	*	*	Whole arm
Use a key to unlock the door+			*	*			*	Whole arm
Comb hair			*	*	*	*	*	Whole arm
Pick up a cup by the handle			*	*		*		Distal
Write on paper					*	*		Distal
Carry an object in the hand			*	*				Whole arm
Dial a number on the phone				*	*		*	Distal
Open a jar						*	*	Distal
Pick up phone					*			Whole arm
Put on T-shirt							*	Whole arm
Do or undo buttons on clothing				*			*	Distal
Do or undo a zip				*			*	Distal
Drink from cup/mug							*	Whole arm
Wash your back							*	Whole arm

Items in bold indicate those retained at the end of round three of item reduction – modified Delphi consensus; Item marked with a + was initially excluded.

During round one, a passive function item, 'Cleaning around the elbow' was removed. This item was removed on the recommendation of eight members of the consultation group (80%), because it was identified as not being relevant for many patients. However a misconception about the meaning of this item may have occurred, which was not immediately evident. Clinicians may not have understood this item to be referring to the elbow crease as well as the extensor surface of the elbow. Clinically this item seems important for patients with flexor spasticity at the elbow. However, its wording may have confused clinicians in the consensus process and would also be likely to confuse patients.

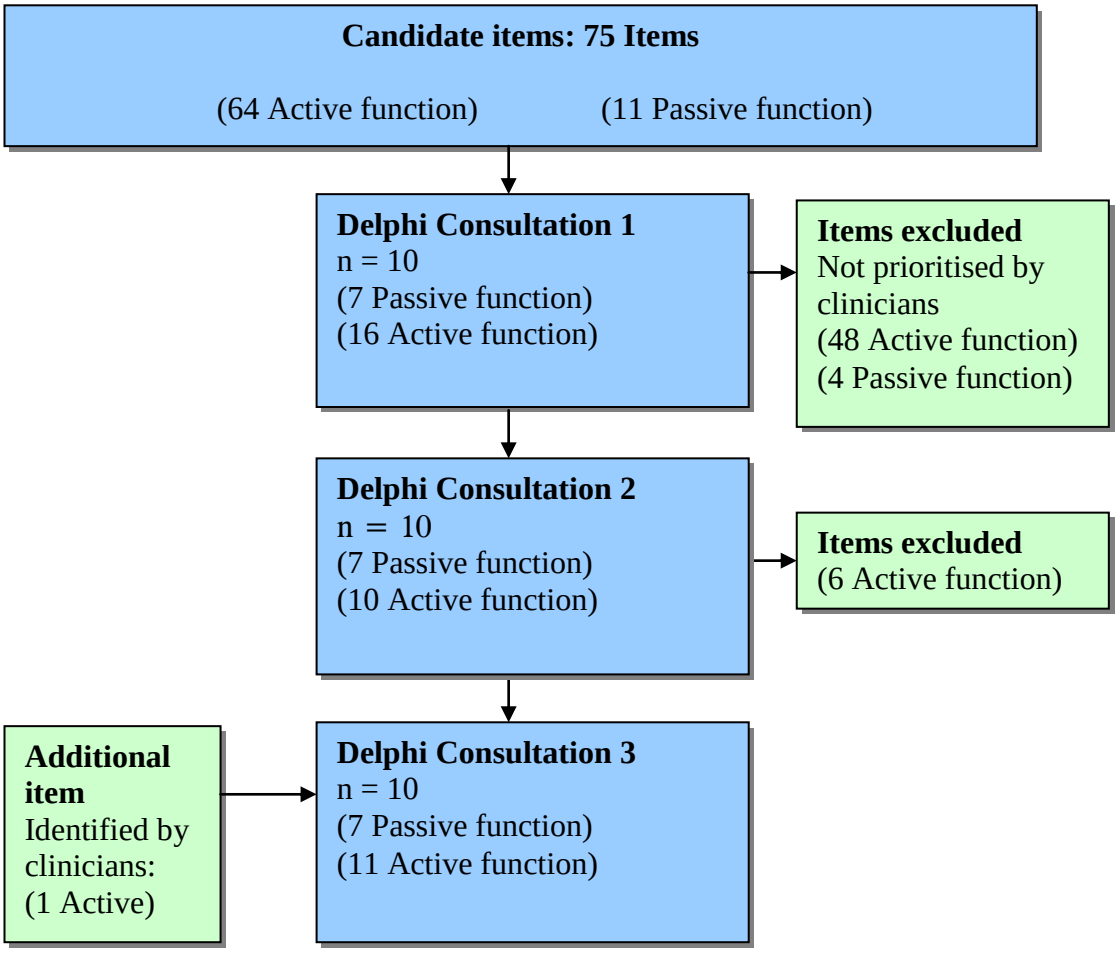
Delphi Consultation Round 2

All 10 clinicians again returned the round two consultation document. A further six active function items were removed following round two. Consensus was between 60% and 80% (6-8 clinicians) for removal of these items. Items not in bold in Table 6.1 were removed.

Delphi Consultation Round 3

All 10 clinicians returned the final round three-consultation document. No further items were excluded and there was between 80 and 100% (8-10 clinicians) consensus for the inclusion of the items chosen. One item which had initially been removed; 'use a key to unlock the door' was re-inserted with the agreement of 80% (8/10 of clinicians (see Table 6.1, item marked with '+'). Figure 6.2 shows a summary of the Delphi consultation process.

Figure 6.2 Delphi consultation - item reduction



6.4.2 Stage 2 - Item confirmation

A total of 58 questionnaires were sent to clinicians and 36 (62%) were returned. The characteristics of these 36 clinicians are shown in Table 6.2.

Table 6.2 Demographic information for wider clinician consultation (n=36)

Professional Group	Numbers (%)
Physiotherapist	25 (69%)
Occupational Therapist	6 (17%)
Nurse	5 (14%)

Thirty-two questionnaires were posted or directly presented to 16 patients and 16 carers. Thirteen questionnaires were completed in each group (81%). Table 6.3 displays the characteristics of the patients and carers returning questionnaires.

Table 6.3 Demographic information of patients (n=13) and carers (n=13)

	Characteristics	Patients	Carers
Age of patients (years)	**Median (range)**	48.5 (30-64)	-
Gender	**Male**	8 (62%)	-
	Female	5 (38%)	-
Ethnicity	**White**	10(77%)	-
	Black	1 (8%)	-
	Asian	2 (15%)	-
Primary Pathology	**Haemorrhagic Stroke**	5 (38%)	-
	Ischemic Stroke	8 (62%)	-
Questionnaire completion method	**Face to face**	8 (62%)	3 (23%)
	Postal Return	4 (31%)	7 (54%)
	Telephone	1 (8%)	3 (23%)

The median and range of ages presented in Table 6.3 is at the younger range of ages in the general stroke population. Recommendations by clinicians, patients and carers (respondents) for the exclusion and inclusion of items following item confirmation are presented in Table 6.4.

Table 6.4 Respondents recommendations for item confirmation.

Items	Respondents recommending removal	Respondents recommending insertion
PASSIVE FUNCTION		
1. Splint application	**1**	-
2. Positioning the arm comfortably	**1**	-
3. Putting on a glove	**2**	-
4. Cutting fingernails	-	-
5. Cleaning the armpit	-	-
6. Cleaning the palm	-	-
7. Putting arm through coat sleeve or dressing the arm	-	-
ACTIVE FUNCTION		
8. Eat with a knife and fork	-	-
9. Pick up a glass, bottle, or can	**1**	-
10. Brush teeth	**1**	-
11. Use a key to unlock the door	**2**	-
12. Comb hair	**1**	-
13. Write on paper	**1**	-
14. Dial a number on the phone	**1**	-
15. Open jar	-	-
16. Do or undo buttons on clothing	-	-
17. Drink from cup/mug	**1**	-
18. Wash your back	**5**	-
19. Tuck in a shirt	-	**4**
20. Effect of the arm on balance when walking	-	**6**
21. Hold an object still while using the unaffected hand	-	**7**
22. Rolling over in bed because of arm tightness	-	**2**
23. Shaving / make-up application	-	**1**

- No preference for removal or insertion

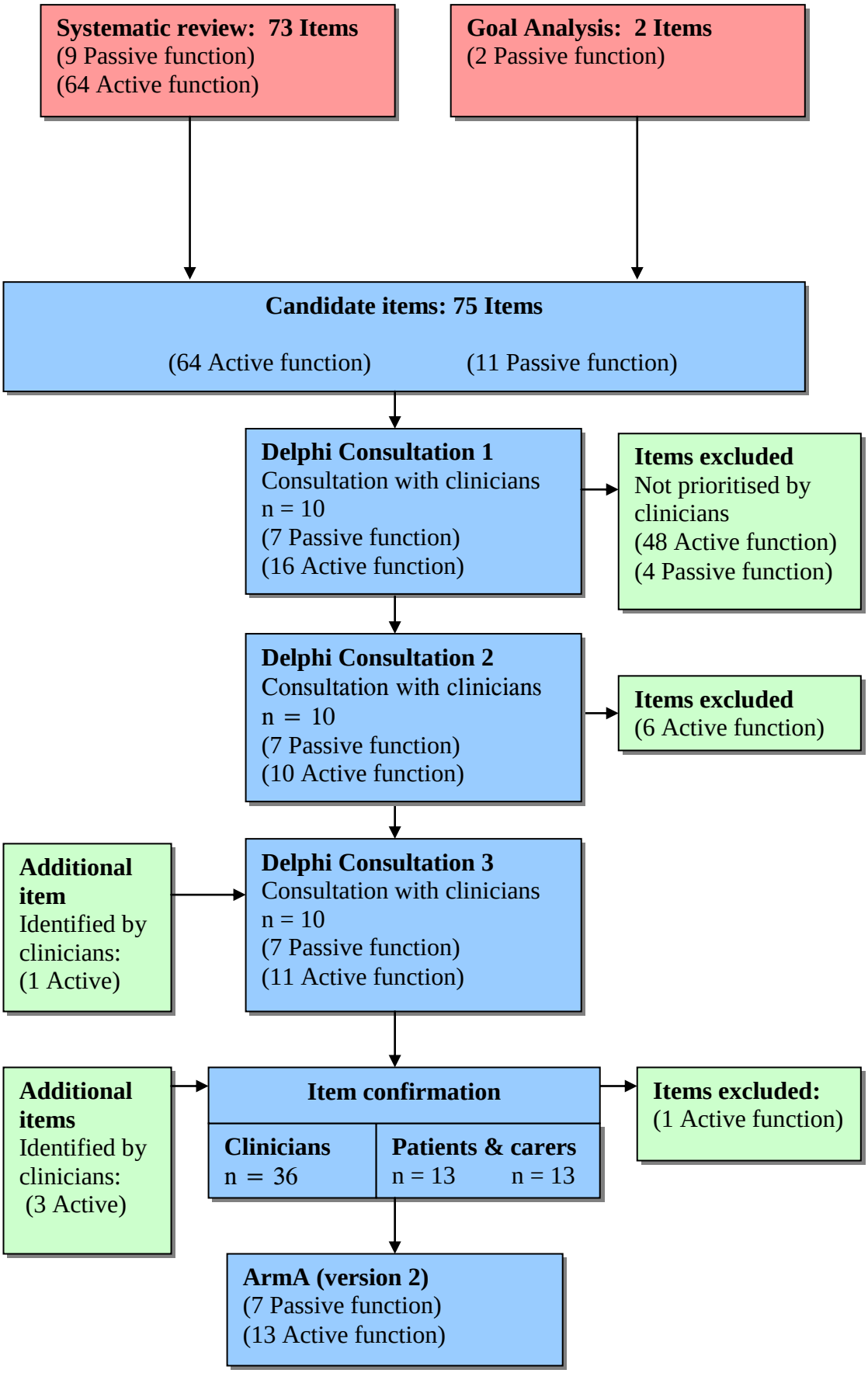
The majority of items were not considered by respondents for removal (n=12), of the other items only one had more than two votes. Five items from the additional list provided, were recommended for inclusion. The specific modifications and the items changed are detailed below.

Modifications

Several modifications resulted from the wider consultation with clinicians, patients and carers. The active function item 'Wash your back' was removed and replaced by 'Tucking in a shirt', since five of the respondents identified that washing your back is done by many able bodied people using an aid, which concurred with views expressed by clinicians during item reduction. Two additional items were added. The 'Effect of the affected arm on balance when walking' was added following comment by six respondents. Two clinicians considered this item to potentially fit in either passive or active function, since although walking is active, the effect of the arm is passive. However, the other four respondents felt it should be in the active function sub-scale. The task 'Hold an object still while using the unaffected hand' was also added following support from seven respondents.

The term 'Within the last week' was replaced with 'In the last seven days'. The instructions for completion of the two main sections were further refined. The final measure consists of two domains, active and passive function. Passive function contains 7 items. Active function contains 13 items. Figure 6.3 displays a summary of the changes to items through the different stages of development. Version 2 of the ArmA for psychometric evaluation is presented in Appendix 12. The ArmA is designed to be completed in a self-report manner by patients and carers. However as an alternative when supported completion is needed for patient or carer, it may be administered as a structured interview by a clinician.

Figure 6.3 Summary of item reduction for the ArmA



6.4.3 ArmA item mapping onto the ICF

The ArmA (version 2) items were mapped onto ICF chapters and sub-categories to assess their relationship to the ICF classification system. Mapping onto the ICF is presented for passive function in Table 6.5.

Table 6.5 The ArmA passive function items classified by ICF code

ArmA Passive Items	ICF Codes		
	Chapter	**Sub-category 1**	**Sub-category 2**
1. Cleaning the palm	5 - Self care	d510 Washing oneself	d5100 Washing body parts
2. Cutting fingernails	5 - Self care	d520 Caring for body parts	d5203 Caring for fingernails
3. Putting on a glove	5 - Self care	d540 Dressing	d5400 Putting on clothes
4. Cleaning the armpit	5 - Self care	d510 Washing oneself	d5100 Washing body parts
5. Putting arm through a sleeve	5 - Self care	d540 Dressing	d5400 Putting on clothes
6. Putting on a splint	5 - Self care	d520 Caring for body parts	d5208 Caring for body parts specified
7. Positioning the arm comfortably	5 - Self care	d520 Caring for body parts	d5208 Caring for body parts specified

ICF: International classification of functioning disability and health

All items in the passive function subscale are found in the self-care chapter of the ICF. The mapping of the active function sub-scale is shown in Table 6.6.

Table 6.6 The ArmA active function items classified by ICF code

ArmA Active Items	ICF Codes		
	Chapter	**Sub-category 1**	**Sub-category 2**
1. Do up buttons on clothing	4 - Mobility	d440 Fine hand use	d4400 Picking up
2. Pick up a glass, bottle or can	4 - Mobility	d430 Lifting and carrying objects	d4300 Lifting
3. Use a key to unlock the door	4 - Mobility	d445 Hand and arm use	d4402 Manipulating d4453 Turning or twisting hands or arms
4. Write on paper	4 - Mobility	d445 Hand and arm use	d4402 Manipulating d4453 Turning or twisting hands or arms
5. Open a previously opened jar	4 – Mobility	d445 Hand and arm use	d4402 Manipulating d4453 Turning or twisting hands or arms
6. Eat with a knife and fork	4 – Mobility 5 - Self care	d445 Hand and arm use d550 Eating	d4402 Manipulating d4453 Turning or twisting hands or arms
7. Hold an object still while using unaffected hand	4 – Mobility	d445 Hand and arm use	d4402 Manipulating d4453 Turning or twisting hands or arms
8. Difficulty balancing when walking due to your arm	4 - Mobility	d450 Walking	
9. Dial a number on home phone	3-Communication	d360 Using communication devices and techniques	
10. Tuck in your shirt	5 - Self care	d540 Dressing	d5400 Putting on clothes
11. Comb or brush hair	5 - Self care	d520 Caring for body parts	d5208 Caring for body parts specified
12. Brush teeth	5 - Self care	d520 Caring for body parts	d5201 Caring for teeth
13. Drink from cup/mug	4 - Mobility	d430 Lifting and carrying objects	d4300 Lifting (glass from table)

ICF: International classification of functioning disability and health

Items in the active function sub-scale are spread between three chapters; self-care, mobility and communication shown in Table 6.6.

6.5 Discussion

The development of ArmA has culminated in a measure comprising seven passive function items and thirteen active function items. All items in the passive function sub-scale are found in the self-care domain of the ICF, which confirms the care related nature of these items. The items in the active function sub-scale are spread across three domains: self-care, mobility and communication, concurring with the more diverse nature of active arm function.

The process has further confirmed content and face validity of the ArmA in ensuring items are representative of active and passive function, despite the different ICF categories identified for active function. The ICF does not take into account passive function as a concept. The language and structure of the ICF are based on individuals carrying out personal care and other tasks themselves rather than being assisted by a carer. Although items will map onto the ICF, the changed nature of the task in passive function, often involving partnership between patient and carer, is not reflected.

6.5.1 Strengths and limitations

The development of the ArmA has three areas of strength relating to the process of development:

1. Methods used for item selection,
2. Method of item reduction,
   Delphi consultation
3. Methods of item confirmation
   Wider consultation
   Pilot testing

1. Item selection (generation)

Selection was undertaken using items from measures identified during the systematic review and patient selected items. These methods provided an extensive list of items. All but two of these items were selected from the six existing measures identified in the systematic review, indicating that other researchers in the field had also found these

items to be relevant. Because there were relatively few passive function items, a further step (goal analysis from clinical practice) was undertaken to ensure full item coverage. In fact this added only 2 further items. Both these aspects of initial development provide support for the face validity of the items within the measure. The use of the goal setting analysis for the patient identified items ensures that items are likely to be clinically applicable and that patient and carer perspectives have been considered at an early stage in development. The structure and scoring mechanism for the ArmA were taken from those measures identified in the systematic review and used methods that had been well tested in these measures.

2. Item reduction – Delphi Consultation

Initial item reduction involved a modified Delphi consultation with 10 selected clinicians. The process ensured content validity, due to the experience of the clinicians in this area of practice and therefore appropriate reduction of items. The modified Delphi consultation was effective due to the rigorous process applied. However, a more robust consensus process could have been used, possibly including further rounds of consultation and higher thresholds for inclusion or exclusion of items. Many studies using Delphi consultation have applied an 80% level of agreement between participants for inclusion or exclusion of items. However, consensus in round three of consultation was achieved at 80% regardless of the 60% criteria used and an 80% criteria would therefore have been unlikely to change the final outcome.

3. Item confirmation – Wider consultation

The wider consultation with clinicians experienced in spasticity management confirmed the selection of items, and also enabled some modification. This consultation ensured comment was obtained on the presentation of the items and the measure as a whole by a larger group of clinicians, patients and carers. The document was sent to all members of the UK Adult Physiotherapy Spasticity Forum, which is a relevant group of clinicians involved with upper limb rehabilitation and spasticity management. The occupational therapists contacted were approached because they worked with the physiotherapists but were a smaller group and the nurses were contacted through rehabilitation services in one large NHS trust.

Selection of all clinical groups could have been enlarged to ensure a true national survey through approaching the respective professional bodies or special interest groups for physiotherapists, occupational therapists and nurses. Breadth of experience among the clinicians may also have been improved by selection through a professional organisation. This approach would have given more support to the content validity of the measure and may have led to a larger consultation with a more consistent national focus. The clinicians involved in the wider review included individuals from different services and different regions of the United Kingdom but did not provide an even spread across regions of the United Kingdom. The group selected was also biased towards physiotherapists and although this professional group undertake much upper limb assessment and spasticity management, they are certainly not the only profession involved. Although occupational therapists and rehabilitation nurses were involved and patients and carers were involved in pilot testing, involvement more widely in development could have provided a more representative group of professionals and future users. However, given that physiotherapists are commonly working in management of spasticity in the UK the approach taken was adequate and produced comprehensive comments.

During pilot testing of the ArmA, the size of the group of patients and carers used could also have been increased. The group was relatively small (n=26), but it is unclear if increasing this would make a difference to achieving feedback that is more informative. A more representative sample could however have been considered, including patients from other services, to ensure that there were no service or practice specific factors affecting their views. However, these limitations, while important considerations, do not invalidate the pilot testing applied for the ArmA, which was sufficient to enable subsequent psychometric testing (see Chapter 7).

A possible limitation of prioritising the items generated using the Delphi process and wider consultation, is that a set of homogeneous items may be produced. An initial 48 active function items and 4 passive function items were excluded at round 1. This was likely to be because participants were asked to identify the most relevant items for patients undergoing focal spasticity intervention (PT and BTX). However, this may risk losing the uniqueness of the broader range of items important for ensuring a wide range to any hierarchical scale. Homogeneity may be a strength in supporting

unidimensionality (in a single or multiple dimensions), but a group of homogeneous items can result in an inability to differentiate between people at extremes of the scale. This may be more of an issue for active function items, many of which were excluded in the Delphi consultation. However, while this is a theoretical concern, in practice it may be less significant because items selected were focused on lower level active function more likely to change in a group undergoing spasticity intervention, but may limit the application of the ArmA for other focal interventions.

Other approaches to evaluation of the draft ArmA (7 passive and 13 active function items) by patients and carers could have been considered. Such approaches could include structured interviews (Reed and Roskell-Payton 1997; Smith et al. 1997) or focus groups (Kitzinger 1994; Sim and Snell 1996). Structured or semi-structured interviews or focus groups may obtain more detailed and expansive feedback from respondents than asking for written feedback as was the case in this review (Sim and Snell 1996).

6.5.2 Comparison with development of other measures

The Motor Activity Log was developed to measure the outcome of CIMT (Taub et al. 1993; Uswatte et al. 2005). The aim of the MAL as with the ArmA was to measure functional performance outside the clinic environment in individuals following stroke. A number of different versions of the MAL have been developed as described in Chapter 4. Detailed description of the development of the original 14-item scale is limited but items were chosen by the authors on the basis of those thought to change following constraint therapy by the development team (Taub et al. 1993). Item selection and reduction for the ArmA was more structured by incorporating items from other measures identified in a systematic manner from the literature, combined with patient or carer selected items. However the approaches have similarities in attempting to identify items, which have clinical relevance to practice, as well as considering items from the perspective of their contribution to the overall measure.

A different approach using Rasch analysis, was used to develop and validate the ABILHAND, a measure of manual ability in rheumatoid arthritis (Penta et al. 1998). Penta and colleagues initially used a relatively small group (n=18) of patients, which

could be challenged as too few for undertaking Rasch analysis. The measurement items were generated either from existing scales or by the researchers, which resulted in 57 initial items. Limited information is provided about initial item identification from the literature or additional items added and how they were selected. It would be desirable to have a clearer methodology for this as given for the ArmA. The number of items was reduced from 57 to 46, the other 11 items were thought not to be measuring the same construct based on Rasch analysis. Unidimensionality and reliability of the measure were supported. Despite the limitation in the number of participants for item generation, Penta and colleagues used the Rasch method to produce a unidimensional single construct scale measuring arm function. The ABILHAND was then further evaluated in 103 patients with chronic stroke (>6 months post onset), which addressed the limitations of sample size with the original study (Penta et al. 2001). The second study provided an acceptable sample size for preliminary Rasch analysis, although still relatively small. Further evaluation of ArmA using the Rasch method may be useful in due course.

Hudak and colleagues undertook the development of the DASH, an upper extremity outcome measure for application in patients with musculoskeletal problems involving the upper limb (Hudak et al. 1996). The measure was developed in three stages. The first stage involved the generation of items from a literature review of measures, which initially produced 821 candidate items from 13 measures. In stage two, item reduction was undertaken by group consensus through three rounds of consultation resulting in a 78-item questionnaire. In stage three of development, reliability and validity testing resulted in the reduction of items to 30 in total (Davis et al. 1999).

Development of the ArmA took a similar approach to that used for the DASH, except that an additional source was added for item generation, involving patients and carers. The ArmA process also included consultation with a wider group of clinicians to confirm item selection and none of the clinicians involved with item selection were involved with the development of the measure. In contrast, those involved in the DASH consultation group also developed the measure.

6.5.3 The role of the user (patient and carer) in item selection and reduction

The involvement of users has benefits at a number of levels, from appropriate research design to consultation about aspects of measures for application in studies and dissemination of findings. The Department of Health and the National Research Ethics Service have both emphasised the importance of patient involvement in research projects and measurement of health outcome (Department of Health 2009; INVOLVE 2009). However there are costs both financial and in time when involving users in research. These issues will be briefly explored and related to the development of the ArmA.

User involvement in research has been defined as users of services being active partners in the research process rather than just research participants (Hanley et al. 2004). User involvement as a concept has undergone development over the past 10 years with moves from very peripheral input, such as obtaining users views during data collection, to user involvement at every stage of development. Partnership may take place in different phases of the research process (Lacey and MacNamara 2000) or may be present all the way through from conception of the project to dissemination of findings (Jones et al. 2009).

In the ArmA development process, early patient involvement was used to identify measurement items through goal setting, but essentially involved users as research participants. Users were again involved in piloting and commenting on the draft version of the ArmA. These instances of user involvement in the ArmA development did not involve users at all stages, which may represent a weakness in the process from a user involvement perspective. However, strength in this involvement is evident in identifying clinically relevant items in a focused manner particularly supported by using clinical goals for spasticity management.

User involvement in both the ArmA development and the work of Lacey and MacNamara was at a consultation level rather than full integration of users into research question generation and project design now being incorporated in some studies (Jones et al. 2009). Another possibility in ArmA development was the inclusion of users at an earlier stage in commenting on the manner and theoretical conception of measurement. However, the approach taken in the ArmA development has resulted in a measure,

which does incorporate items important to patients and carers as evidenced in the pilot testing. The ArmA development has also benefited from consultation and pilot testing of the measure with patients and carers. Given these aspects of patient and carer input, involvement of users in development is considered adequate.

Patient Reported Outcome Measures (PROMs) such as the ArmA, are receiving attention in the rehabilitation literature (Playford 2008). User involvement in the development of such measures is emphasised in generating items (Thissen et al. 2007; Playford 2008). Tennant has given qualified support to the use of approaches such as GAS, as used in this thesis, in identifying items (Tennant 2007). Items identified in this way can then undergo Rasch analysis to determine the possible scaling properties of the set of items identified and fit to the Rasch model. Alternatively, non-parametric IRT methods such as Mokken analysis may be used to initially consider the ordinality of the data.

6.6 Conclusion

The use of Delphi consultation with the addition of further clinician and patient involvement has resulted in a measure, which should provide important clinical information and be feasible in practice. The process of item selection, reduction and confirmation was comprehensive and while limitations to the methodology are present, the overall process had a high degree of rigour ensuring confidence in content and face validity of the list of items produced.

The next phase of development for the ArmA (version 2) was psychometric testing, which is presented in Chapter 7. Version 2 of the ArmA will be referred to as the 'ArmA' for the remainder of the thesis.

Chapter 7 Evaluation of ArmA properties and application

7.1 Introduction

The purpose of this part of the overall project was to demonstrate the psychometric properties of the ArmA in a patient sample of sufficient size. This psychometric analysis was undertaken in the context of a prospective cohort study of BTX and PT intervention in the clinical setting. In this Chapter, the two inter-linked sub-studies that were carried out are described.

Sub-study 1 – An evaluation of the psychometric properties of the ArmA

The global psychometric concepts explored comprised reliability, validity, unidimesionality, ordinal scaling and responsiveness to change. These psychometric properties were discussed in theoretical terms in Chapter 3 (pages 94-112), and specific methods for evaluation are described in the methods section (pages 201-204 of this chapter). Criteria adapted from Terwee and colleagues (Terwee et al. 2007) were applied to evaluate the psychometric properties of the ArmA (see Table 7.1 for an overview).

Table 7.1 Quality criteria applied in the ArmA evaluation.

Psychometric Requirement	Evaluation Criteria
Construct validity Relate to other measures appropriately.	Correlation with other measures for convergent and divergent validity.
Homogeneity of measurement items The unidimensionality of a set of items.	Preliminary evaluation of unidimensionality using Principal component analysis and Mokken analysis.
Scaling Extent to which measurement items form an ordinal or interval scale	Preliminary evaluation of ordinal scaling properties using Mokken analysis.
Internal consistency Extent to which items in scale correlate.	Evaluation using Cronbach's Alpha: Rating between 0.70-0.95.
Test re-test reliability (Repeatability) Extent to which patients can be distinguished despite measurement error (relative measurement error).	Test re-test at two time points evaluated by weighted Kappa > 0.70.
Responsiveness Detects change over time relating to actual change occurring.	Demonstrating change over time when change is expected. Comparison of responders with non-responders, Effect Size (ES) and Standardised Response Mean (SRM) estimations.
Interpretability The degree to which qualitative meaning can be assigned to quantitative scores.	Preliminary indication of minimal important change (MIC).
Floor and ceiling effects Number of respondents achieving the lowest or highest scores.	Less than or equal to 15% of respondents at either extreme of scale (or sub-scale).
Feasibility Suitable for routine clinical use, while maintaining psychometric properties.	Using a questionnaire with participants, evaluate use of the measure in clinical practice with psychometric properties retained.
Burden The time, effort or other demands of administering the measure.	Using a questionnaire with participants to obtain information on time taken and effort to complete the measure.

The sub-study also enabled the feasibility of the ArmA to be determined. This element was added because the ease with which the ArmA could be used in standard clinical practice environments was considered to be an essential requirement.

Sub-study 2 – A hypothesis-generating cohort investigation of the course of functional changes

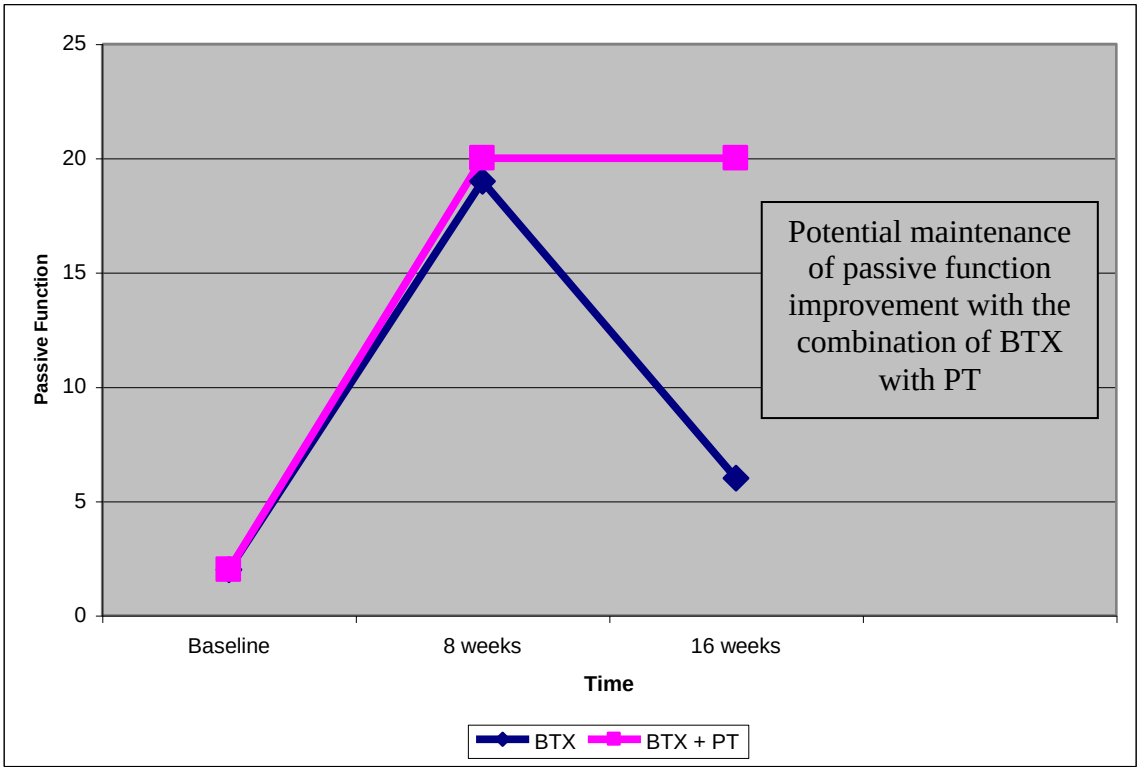
The cohort study was undertaken using the ArmA to explore changes in passive function and spasticity occurring after intervention with BTX and PT interventions at 8 weeks and at 16 weeks.

Reduction in spasticity has been demonstrated in a number of randomised trials following BTX injection (Bhakta et al. 2000a; Bakheit et al. 2001; Brashear et al. 2002; Bakheit et al. 2004b; Bhakta et al. 2008). Change in function has also been demonstrated by some authors for passive function (Bhakta et al. 2000a; Francis et al. 2004). However, the clinical effect of BTX alone would be expected to reduce as BTX is eliminated from the neuromuscular junction from approximately six to eight weeks onward (Brin 1997; Bell and Williams 2003), potentially leading to a decrease in previous functional gains.

Despite this, some studies have also indicated that passive functional improvements can be maintained at least to 12 weeks following administration of BTX (Bhakta et al. 2000a; Brashear et al. 2002). In addition, in the secondary analysis by Francis and colleagues (Francis et al. 2004), functional improvements were found to take longer to develop in a minority of patients than reduction in spasticity. These authors suggest that PT interventions such as positioning or serial casting regimes may be important in reaching and maintaining improvements following BTX (Francis et al. 2004).

In summary, intervention with BTX in the absence of PT would be expected to lead to reduction in spasticity and improvement in passive function by 8 weeks after injection followed by a reversion to near pre-intervention levels by 16 weeks as the effect of BTX wears off (see Figure 7.1). In contrast, when applying PT in combination with BTX it is possible that the improved level of passive function will be maintained to 16 weeks post injection (see Figure 7.1).

Figure 7.1 A diagrammatic representation of the potential for maintenance of - passive function following BTX and PT



The aim of the cohort study was to test this model of passive function maintenance in the presence of increasing spasticity using the ArmA. The study also aimed to include systematic documentation of which PT interventions had been applied and in what manner in each case. It was anticipated that this would also lead to the generation of clinically important research questions for future evaluation.

7.2 Objectives

This chapter describes the method for addressing the remaining three objectives.

Sub-study 1 – Psychometric evaluation

Objective 5. To evaluate the reliability, internal consistency, construct validity, unidimensionality and ordinal scaling of the Arm Activity measure (ArmA) – a measure of difficulty in active and passive function.

Objective 6. To evaluate the responsiveness and feasibility of the measure when using it to assess outcome following spasticity intervention in the upper limb.

Sub-study 2 – Cohort study

Objective 7. To apply the ArmA in measuring change in passive and active function following spasticity management intervention.

In this cohort study, the utility of ArmA will be tested through the following null hypothesis. The hypothesis addresses passive function, but not active function because change in this sub-scale was expected in only a minority of participants undergoing spasticity management.

Null hypothesis: *Improvement in passive function measured by the ArmA at 8 weeks following BTX and PT intervention will not be maintained above baseline levels as the effect of the BTX on spasticity decreases.*

7.3 Methods

Design

Psychometric evaluation of the ArmA was undertaken on the passive and active function sub-scales of the measure. Two groups of participants were recruited for the evaluation.

- Group 1, a prospective consecutive sample undergoing spasticity intervention were the primary group used for evaluation of the ArmA measure psychometric properties.

- Group 2 were purposively selected after it became apparent that the large majority of subjects in group 1 had no active function, resulting in the ArmA scores not being representative of the full range of the scale. In addition, a greater sample size was required to fulfil requirements for psychometric evaluation. Group 2 were a more able group selected on the basis that they were able to perform at least one of the active function activities within the ArmA. This group were not undergoing management for spasticity.

Both groups were assessed at baseline (Time 1) and one day later (Time 2). In addition, group 1 were assessed at 8 weeks (Time 3) and 16 weeks (Time 4).

Sub-study 1 – Psychometric evaluation of the ArmA

Data collected at baseline (Time 1) and one day later (Time 2) were used to evaluate reliability (repeatability). Data collected at baseline (Time 1) was also used in evaluation of construct validity, internal consistency and to confirm the dimensions of the active and passive function sub-scales using Principal Components Analysis and Mokken analysis. Mokken analysis was also used to undertake a preliminary evaluation of ordinal scaling. Data at 8 weeks (Time 3) were used following intervention to evaluate responsiveness.

Sub-study 2 – Functional change evaluation: **A cohort study**

Data at baseline (Time 1), 8 weeks (Time 3) and 16 weeks (Time 4) were used for comparison of change following intervention in spasticity and passive function.

Setting

Data were collected at two sites during the psychometric evaluation and cohort study to ensure recruitment of sufficient participants in groups 1 and 2. The sites were the Regional Rehabilitation Unit (RRU), Northwick Park Hospital and the Alderbourne Rehabilitation Unit (ARU), Hillingdon Hospital.

Both services provide inpatient rehabilitation, outpatient clinics including spasticity clinics and, in the case of the RRU, an outreach service providing specialist spasticity management intervention. Participants at both sites were managed through the spasticity management ICP developed on the Regional Rehabilitation Unit, Northwick Park Hospital (see Appendix 11 for current version).

Sample

The sample size was based on the criteria by Terwee and colleagues for evaluation of construct validity and test re-test reliability in groups of at least 50 participants (Terwee et al. 2007). However, given that unidimensionality is an important requirement for measurement, this was considered with relation to the sample size for this work.

Determination of sample size can be based on two broad methods. 1. Subject to variable ratios have been proposed (Pedhazur 1997), 2. alternatively total sample sizes for the study group have also been used (Aleamoni 1976; Barrett and Kline 1981; Comfrey and Lee 1992).

1. Authors have made a number of different recommendations using the ratio method. When undertaking PCA, Nunnally and Bernstein (page 102) recommended a sample size with a ratio of 10:1 subjects to items (Nunnally and Bernstein 1994) and Terwee and colleagues (2007) a ratio of 7:1 with a total sample of 100 or greater. The ArmA has 20 items, so using these ratios results in a sample of 200 or 140 respectively. Gorsuch (1983; p. 332) and Hatcher (1994, p 73) both made recommendation for a minimum subject to item ratio of 5:1 in exploratory factor analysis (EFA), which would result in a sample of 100 for the current work.

2. A range of sample sizes have also been suggested for minimal total sample size, ranging from: 50 giving very poor adequacy to 1000 giving excellent adequacy of sample size according to Comfrey and Lee (1992). Barrett and Kline (1981) have also recommended sample size with a minimum of 50, while Aleamoni (1976) suggests a minimum of 400.

Briefly, considering Mokken analysis (also discussed in Chapter 3; application of Mokken analysis, page 104), Molenaar (2000) recommends a lower limit of 100 in Mokken analysis for total sample size. The studies by Van der Lee and colleagues (2002) and van der Putten and colleagues (2005) had participant numbers as low as 66 and 63 respectively. DeJong and Molenaar also used a sample of 82 in an example of Mokken analysis application (DeJong and Molenaar 1987).

In summary, a review of studies applying EFA or PCA concluded, that absolute minimum sample sizes, rather than subject to item ratios, are the most relevant (Guadagnoli and Velicer 1988). In contrast, many advocates of the ratio of items to subjects, suggest that this is the superior method (Gorsuch 1983; Hatcher 1994; Osborne and Costello 2004). There is therefore no universally agreed approach to identifying required sample sizes for these methods. Absolute sample sizes however seem over simplistic, given as suggested by Osborne and Costello (2004), the variance in the types of scales to be examined and the variation in the number of items in such scales. Nevertheless Osborne and Costello (2004) consider both the ratio and the total minimum number of participants to be important in undertaking PCA. Consideration of item to subject ratio has however, been considered of prime importance in this thesis.

The aim in this thesis was therefore, to recruit participant numbers of 100, to ensure robust preliminary findings consistent with the recommendations of Gorsuch (1983), Hatcher (1994) and Molenaar (2000). In addition, this conforms to the recommendations of Barrett and Kline (1981) for minimal total sample size of 50 participants or more. In common with other psychometric methods used in this thesis the application of PCA and Mokken analysis, represent preliminary evaluations of the ArmA measure.

Inclusion and exclusion criteria:

Inclusion criteria were:

- Hemiplegic upper limb impairment affecting either active or passive function.
- Age between 18 and 85 years.

In addition for Group 1

- Undergoing treatment for spasticity management in the upper limb requiring BTX intervention and PT.

Reasons for exclusion were:

- Patient declines to participate or family and/or treating team declines on their behalf.
- Unable to complete a questionnaire and no carer (professional or family) available to undertake questionnaire completion. Examples of situations that may lead to exclusion are indicated below:
    - Does not speak English
    - Unable to communicate responses (where feasible communication using adapted methods will be employed, including support of a speech and language therapist to avoid exclusion).
    - Does not have the cognitive ability to understand the questions.

Measures

Eight measures were used.

Arm Activity Measure (ArmA)

The ArmA (shown in Appendix 12) is a measure of difficulty in passive and active arm function. The ArmA comprises a seven-item passive function subscale and thirteen-item active function subscale. It has a Likert scoring system between 0 (No difficulty) and 4 (unable to do task). It is self-rated by the patient or carer for a period of the preceding 7 days. If activities have not been performed in the past 7 days but are possible, then a 'best estimate' was made of the task. For passive function, if a carer was involved, then they agree the score with the patient as both contribute to completion of the activity. If only the carer was involved in completion of the activity, then the carer alone rates difficulty for that item. Active function is scored by the patient. If the patient has no active function ability, this is recorded as being 'unable to do the tasks' even if the carer is undertaking scoring. A total of the individual item scores are made for both sub-scales, but sub-scale totals are not combined. The passive function sub-scale scores range from 0-28 and the active function sub-scale scores range from 0-52.

Leeds Adult Spasticity Impact Scale (LASIS)

The LASIS (Bhakta et al. 1996; Bhakta et al. 2000a) (Shown in Appendix 13) is a measure of the impact of spasticity on arm function, for use in evaluating spasticity management intervention. The LASIS has two sub-scales: a disability sub-scale consisting of 12 items and a carer burden sub-scale consisting of 9 of the same items. The LASIS uses a scale between 0 (No difficulty) and 4 (Unable to do task). The patient and carer rated the LASIS over the preceding 7 days. The LASIS was either completed as a structured interview or self-completed by the patient and carer on a small number of occasions. A modified approach to scoring was used. When patients and carers were both involved in the activity, the two scores were combined to produce a mean. Items 1 to 9 were classified as passive function and items 10 to 12 were classified as active function to allow comparison with the ArmA sub-scales.

Disabilities of the Arm Shoulder and Hand (DASH)

The DASH (Hudak et al. 1996) (Shown in Appendix 14) questionnaire comprises 30 items, 21 of which are arm active function items, 5 are symptom related, for example pain and the remaining 4 items review the impact of arm impairment on well being and participation (there are no passive function items in the DASH). The DASH uses a scale between 0- none and 5- extreme difficulty for functional tasks. For other domains the six point scale is still used, but adapted with relevant descriptors. The patient rated the DASH over the preceding 7 days. If activities have not been performed in the past week but were possible, then a 'best estimate' was made on the task. The DASH is designed for self-completion in musculoskeletal upper limb impairment, but was applied in a neurorehabilitation group for this study. The DASH items are usually summed into one total, which is then divided by the number of responses to the 30 questions (if more than three responses are missing a total is not generated), 1 is then subtracted and the total multiplied by 25. In this study for comparison with the ArmA an overall total for the measure was not produced in this way, instead a total sum was produced for the active function items only (DASH Active; items 1 to 21) for comparison with ArmA.

Goal Attainment Scaling (GAS)

GAS (Ashford and Turner-Stokes 2006) (Shown in Appendix 3) is a method of scoring the extent to which patient's individual goals are achieved in the course of intervention. In GAS, goals are individually identified to suit the patient, and the outcome levels set around their current and expected levels of performance using the standard method. The GAS process provides a consistent framework for recording goals. The GAS is scored on a 5-point scale, with anchor points for attainment at each level set before intervention. Scoring of GAS followed the approach proposed by Turner-Stokes (Turner-Stokes 2009b).

If the patient achieves the expected level, this is scored at 0.

If they achieve a ***better*** than expected outcome this is scored at:

+1 (Somewhat better)

+2 (much better)

If they achieve a ***worse*** than expected outcome this is scored at:

-1 (Somewhat worse) or

-2 (much worse)

An important part of GAS is the establishment of the clinical outcome which is considered as successful before the start of intervention (see Appendix 3). If the patient was unable to contribute to goal setting, goals were agreed by the carer and treating clinician. For each goal, the spasticity clinic team, dependent on achievement of the predefined outcome, assigned an attainment level. A full description of the GAS procedure is provided in Appendix 3.

Clinician's Categorisation of Response (CCR)

The Clinician's Categorisation of Response was the treating clinician's rating of the overall outcome following intervention. The treating physiotherapist or occupational therapist (not part of the spasticity clinic team) was asked to categorise the outcome of the intervention as either, 'responder' or 'non-responder' when reflecting on the goals of intervention set, but separately and without knowledge of final GAS scores (Brashear et al. 2002).

Barthel Index (BI)

The Barthel Index (Wade and Collin 1988) (Shown in Appendix 15) is a measure of global disability and function. The Barthel Index self-completion version was completed by patient or carer (Gompertz et al. 1994). The measure comprises 10 items relating to personal ADLs, with each item scored on either a scale of 0 to 2 or 0 to 3. The total scale ranges from 0 (total dependence) to 20 (complete independence) in the version used in this study.

Feasibility Questionnaire (FQ)

The feasibility questionnaire (Shown Appendix 16) was used to evaluate ease of use, relevance and value in the clinical situation. It comprises one question each for time to complete, relevance, usefulness of the active function section, usefulness of the passive function section and ease of completion. The FQ was designed for completion by patients and carers for evaluation of ArmA. Each question is rated on a 5 point Likert scale.

Modified Ashworth Scale (MAS)

The MAS (Wade 1992b; Brashear et al. 2002) (Shown Appendix 17) is a clinical measure of spasticity, which is widely used in research (see Table 1.3; page 44) and clinical practice. The MAS forms a single item scale from 0 (no increase in muscle tone) to 4 (affected part rigid in flexion or extension), with an additional point at +1 (slight increase in muscle tone…) producing a six-point scale (see Appendix 17). The MAS therefore provides a single score to represent spasticity. The MAS is scored by grading the resistance to passive stretch and was scored in this study by the spasticity clinic team. A single score was given for measurement at each joint at each time point. If the MAS was recorded for more than one joint, the mean was used to represent total spasticity affecting the upper limb for comparison to functional change on other measures.

7.3.1 Procedure

Identification

Group one (G1)
All patients referred to the spasticity services at the RRU and ARU over a 21 month time period were offered the opportunity to participate. Participants and carers were initially approached by their treating physiotherapist or occupational therapist and asked to participate.

Group two (G2)
Patients were offered the opportunity to participate by their treating physiotherapist or occupational therapist.

All patients were given an information sheet about the project (see Appendix 8). The researcher answered any questions they might have before gaining consent (see Appendix 9).

Recruitment

In cases where the participant was unable to sign (for example in the case of impaired arm function), a witness signed to indicate consent had been given. In cases where patients were unable to consent due to cognitive impairment assent confirmed in writing was sought and obtained from the next of kin and treating team.

Data collection

Baseline (Time 1) assessments comprised all 8 measures. 4 measures (ArmA, DASH, BI and FQ) were self-completed by the person carrying out the care activity (patient and/or carer). Of the remaining 4 measures, MAS and CCR were completed by the spasticity clinic team and the LASIS was completed as a structured interview. The GAS goals were set by the patient and/or carer in conjunction with the spasticity clinicians. The clinicians formally documented the identified GAS goals.

All measures were completed at the clinic appointment where possible. The order of presentation of the measures was counter-balanced to limit bias from order effects. Patients and carers unable to complete measures at the clinic appointment took the questionnaires away and returned them in a 'freepost' envelope. This occurred in a minority of cases when for example they were dependent on hospital transport.

One day follow-up (Time 2) the ArmA was returned by post for outpatients or outreach-patients and collected by hand for in-patients.

Eight week (Time 3) and 16 week (Time 4) assessments were undertaken in the clinic or at a prearranged appointment. Table 7.2 indicates the measures completed at each time point.

Table 7.2 Measures completed at each time point

Measures	Time 1 Baseline	Time 2 Baseline plus One day	Time 3 Baseline plus 8 weeks	Time 4 Baseline plus 16 weeks
	Groups 1 and 2	**Groups 1 and 2**	**Group 1**	**Group 1**
ArmA	√	√	√	√
LASIS	√	-	√	√
DASH	√	-	√	√
GAS	√	-	√	√
Clinicians Categorisation of Response	-	-	√	√
Feasibility Questionnaire	√	-	-	-
Barthel Index	√	-	√	√
Modified Ashworth	√	-	√	√

Data collected = √

See Section 7.4.1 and Table 7.3 (page 205) for the number of participants approached, recruited and reviewed at each time point.

Intervention

Participants received BTX administration following baseline (Time 1) assessment as part of normal clinical practice. This was followed by PT interventions appropriate to clinical need over the subsequent 16 weeks. The treating therapist(s), usually following recommendation by the spasticity clinic team, gave concomitant intervention (e.g. splinting of the wrist and hand). The BTX and PT interventions are described in the results of the cohort study.

Data Management and Error Checking

A list of patients recruited and their stage of progression through the study was recorded electronically in a password-protected file. Paper records for individual patients were stored in a locked filing cabinet and electronic records were anonymised and password protected.

The researcher and a colleague undertook double entry of all the data apart from Modified Ashworth Scores to ensure accuracy of insertion into the database for

analysis. Modified Ashworth Scores were entered by the researcher and checked by another researcher in 20 cases. For double entered data, these were entered into two duplicate data entry sheets without reference to each other. The two versions were then compared and errors identified. When errors occurred these were checked and corrected by referring to the original paper questionnaires.

7.3.2 Analysis – Psychometric methods

Data analysis was undertaken using SPSS v15 (SPSS 2000) and STATA v10 (Stata 2001) statistical analysis packages. Mokken analysis was undertaken using MSPWIN 5.0 software package (Molenaar et al. 2000). The evaluation of psychometric properties is referred to in Chapter 3 (Section 3.3.3; page 94 and Section 3.5; page 113), the methods of evaluation for these properties are referred to again here specifically as applied in this chapter.

Floor and ceiling effects

Floor and ceiling effects are normally examined to determine if items are missing, in terms of the range of the scale, at the extremes of the measure. However, in the case of the ArmA, floor effects for difficulty were anticipated in the passive function sub-scale in participants who had high active function ability. It was also expected that participants who had high passive function difficulty, would have significant active function difficulty. Floor and ceiling effects were assessed in the study population by considering the percentage of participants at either extreme of the subscales according to the criteria by Terwee and colleagues (Terwee et al. 2007).

Construct validity

Construct validity was evaluated by comparing the ArmA with components of the LASIS and the DASH. The LASIS and DASH were divided into their active and passive function items and the relevant items were correlated with the relevant sub-scale of the ArmA. All scales used are ordinal and therefore do not meet the criteria for the application of parametric tests. Comparisons were therefore undertaken using the Spearman rank order correlation coefficient.

Convergent validity was tested in the passive function sub-scale of the ArmA by comparing it with the LASIS passive items and a high positive correlation was expected. Convergent validity was tested in the active function sub-scale of the ArmA by comparing it with the LASIS and DASH active function items. A high positive correlation was expected with both.

Divergent validity was evaluated by comparing the ArmA passive function sub-scale with the LASIS and DASH active function items. No significant correlation was expected. Divergent validity was tested in the active function sub-scale of the ArmA by comparing it with the LASIS passive function items, with no significant correlation expected.

Unidimensionality and scaling

The dimensionality of the ArmA sub-scales were initially evaluated using principal component analysis. The results from principal component analysis were evaluated by initially considering Eigenvalues above 1 according to the criteria by Kaiser (1960). Following evaluation of Eigenvalues, Scree plots were then examined to support the findings. However to confirm these findings and to provide more objective criteria for the acceptance of the components identified, a Monte Carlo analysis was carried out according to the method by Horn (1965).

Mokken analysis (monotone homogeneity) was also applied to confirm the constructs of the ArmA (see Chapter 3). Mokken analysis was then applied in a preliminary evaluation of the ordinal structure of the items in the ArmA sub-scales.

Internal consistency

Cronbach's alpha was used to evaluate internal consistency applying the criteria of Terwee and colleagues (Terwee et al. 2007). A positive rating for internal consistency was given when ratings for Cronbach's alpha were between 0.70 and 0.95 (Terwee et al. 2007).

Reproducibility

Reproducibility of ArmA was evaluated using Quadratic Weighted Kappa coefficient for test re-test reliability. Time 1 (Baseline) data were compared with data recorded at Time 2 (one day later).

Responsiveness

Responsiveness of the ArmA was evaluated between Time 1 (baseline) and Time 3 (8 weeks) following BTX injection. Responsiveness was determined by comparing the ArmA detection of functional improvement with the Clinicians Categorisation of Response (CCR) in the passive and active function sub-scales. Change in the ArmA at Time 3 (8 weeks) was compared between responder and non-responder categories using the non-parametric Mann-Whitney U test. It was expected that the ArmA would identify a significant difference between the responder and the non-responder groups for passive function as defined by the CCR at Time 3 (8 weeks) following baseline. Responsiveness was further evaluated between responders and non-responders at Time 4 (16 weeks) to further support the findings and as initial evidence of longitudinal validity.

To enable comparison with other measures of upper limb function, effect size and standard response mean were calculated for the ArmA sub-scales, LASIS active and passive items, Barthel Index and DASH active items. Effect size and standard response mean are commonly used with ordinal scales despite being parametric techniques (also see Chapter 3, Section 3.3.3.7; page 109). Due to the acknowledged limitations of this parametric approach, positive and negative rank differences are also presented for each measure from baseline to 8 weeks.

Interpretability

Minimal Important Change (MIC) was calculated using two methods; a criterion-based method and a distribution-based method (also see Chapter 3, Section 3.3.3.8; page 111). The criterion-based method was produced by calculating the mean change in ArmA passive and active sub-scales in the responder group. The distribution-based method as recommended by (Norman et al. 2003) was calculated by using half the baseline (Time 1) standard deviation for ArmA as an estimate of MIC. The calculation of MIC again

uses parametric assumptions and therefore can only provide a preliminary indication of interpretability for the ArmA. Re-evaluation of MIC will be needed once Rasch analysis has been applied to establish the interval scaling properties of the ArmA using a logit scale. Preliminary predictions of MIC were further examined by calculation of sensitivity and specificity for a 1-point change in ArmA, a 2-point change and a 3-point change according to the classification of response by CCR.

Feasibility

Feasibility is concerned with ensuring that outcome measures can a) be practical to use in routine practice and b) retain their psychometric properties, thus ensuring utility (Slade 2002b). Feasibility is evaluated using a self-completed questionnaire administered following ArmA completion. Patients and carers rated the ease of use, relevance, and value in the clinical situation of the ArmA.

7.3.3 Analysis – Evaluation of functional change: A cohort study

Evaluating spasticity intervention using the ArmA

In order to compare change between baseline and both outcome evaluation points the Friedman test was applied (Altman 1991) (p. 334-6). The Friedman test is a non-parametric test of significance for three or more conditions for same or matched subject designs. The Friedman test can only be used to identify general differences between groups, and does not indicate if one group has improved compared to the other. The test is therefore always two tailed. The Wilcoxon test was then applied following the Friedman test as per the recommendations of Altman (Altman 1991) (p. 203-5). The Wilcoxon test was used to compare Baseline to Time 3 (8 weeks), Baseline to Time 4 (16 weeks) and Time 3 (8 weeks) to Time 4 (16 weeks) separately.

7.4 Results - Psychometric evaluation

7.4.1 Demographics

A total of 103 patients were screened to participate in the psychometric evaluation (63 Group 1, 40 Group 2). Of those patients approached in Group 1, four did not have BTX intervention after full assessment and were excluded and one declined to participate. In Group 2, six of the 40 patients initially approached declined to participate.

A total of 92 patients were recruited at baseline, 58 in Group 1 and 34 in Group 2. Table 7.3 shows the response rates at each time point.

Table 7.3 Response rate at each time point

Time Point	Group 1	Group 2	Combined
T1 – Start of baseline assessment (day 0)	58 (100%)	34 (100%)	92 (100%)
T2 – 1 Day following baseline	44 (76%)	34 (100%)	78 (85%)
T3 – 8 Weeks follow-up	53 (91%)	n/a	n/a
T4 - 16 weeks follow-up	48 (83%)	n/a	n/a

n/a = Not applicable

The study sample is described in Table 7.4.

Table 7.4 Demographic characteristics of the study population (n=92)

Groups	Group 1 (n=58)	Group 2 (n=34)	Combined (n=92)
Mean age (years)	47 (SD=17.5)	42 (SD=15.8)	44.5 (SD=16.7)
Male/female ratio	32:26	22:12	54:38
DIAGNOSIS			
Stroke	30 (52%)	18 (52%)	48 (52%)
Right hemisphere	13 (22%)	10 (28%)	23 (25%)
Left hemisphere	17 (30%)	8 (24%)	25 (27%)
Acquired brain injury	22 (38%)	6 (19%)	28 (31 %)
Traumatic	16 (28%)	5 (15%)	21 (23%)
Anoxic	6 (10%)	1 (3%)	7 (8%)
Other	6 (10%)	10 (29%)	16 (17 %)
Multiple Sclerosis	4 (6%)	2 (6 %)	6 (7%)
Motor neurone disease	1 (2%)		1 (1%)
Encephalitis	1 (2%)		1 (1%)
CNS Tumour		4 (11%)	4 (4%)
Spinal cord injury		2 (6%)	2 (2%)
Vasculitus		1 (3%)	1 (1%)
Critical care neuropathy		1 (3%)	1 (1%)

Group 1 participants had a median Barthel Index score of 5.0, (inter-quartile range 0-15) with a high level of disability and dependence in ADL. Group 2 participants had a median Barthel Index score of 14, (inter-quartile range 10-15) with a lower level of disability and dependence.

Entry of data

Double entry of all data (excluding MAS for which a sample of 20 participants was used) identified minimal errors between the two data sets, which are summarised in Table 7.5.

Table 7.5 Errors found and corrected following double entry for each measure

Time point	ArmA	LASIS	DASH	GAS	CCR	MAS
Number of errors	34 (0.5%)	71 (3.1%)	8 (0.1%)	3 (1.6%)	2 (1.6%)	0 (0%)

Overall, data entry error was 3.3% before correction. Errors were resolved by referring back to the original paper copies of the data.

Missing data

Table 7.6 displays the return rate for ArmA, LASIS, DASH, GAS and CCR at each time point.

Table 7.6 Return rate for ArmA, LASIS, DASH, GAS and CCR

Measure	Time 1 Baseline	Time 2 1 day	Time 3 8 weeks	Time 4 16 weeks
Groups 1 and 2 (n=92)				
ArmA	92 (100%)	78 (85%)	-	-
Group 1 only (n=58)				
ArmA	58 (100%)	44 (76%)	51 (88%)	39 (67%)
LASIS	57 (98%)	-	51 (88%)	39 (67%)
DASH	58 (100%)	-	50 (86%)	37 (64%)
GAS	58 (100%)	-	53 (91%)	47 (81%)
CCR	-	-	52 (90%)	46 (79%)

The missing ArmA measure data at Time 2 were due to postal questionnaires not being returned by participants. Missing data in returned measures was 1.4%. The chi-square test was applied at all time points to evaluate if demographic differences could explain the reason for missing data but none were found. Imputation methods were considered but not applied because of the risk of over emphasising the relationship between data points for psychometric purposes. Case-by-case exclusion of data for each analysis was applied.

Completion of the ArmA

Table 7.7 presents the person who completed the ArmA; patient (professional assistance is also indicated), carer or combined (both patient and carer undertaking completion together).

Table 7.7 Person completing the ArmA (patient, carer or combined)

Completion	Time 1 Baseline	Time 2 1 day	Time 3 8 weeks	Time 4 16 weeks
Patient alone (supported by a professional)	53 (11)	52 (5)	21 (8)	20 (6)
Carer alone	31	19	25	20
Combined patient and carer completion	8	7	7	7

Descriptive statistics

Psychometric evaluation of the ArmA incorporated comparison with a number of different measures. The descriptive statistics for these measures or sub-sections of those measures are shown in Table 7.8.

Table 7.8 Descriptive statistics (median and inter-quartile range) for the study measures.

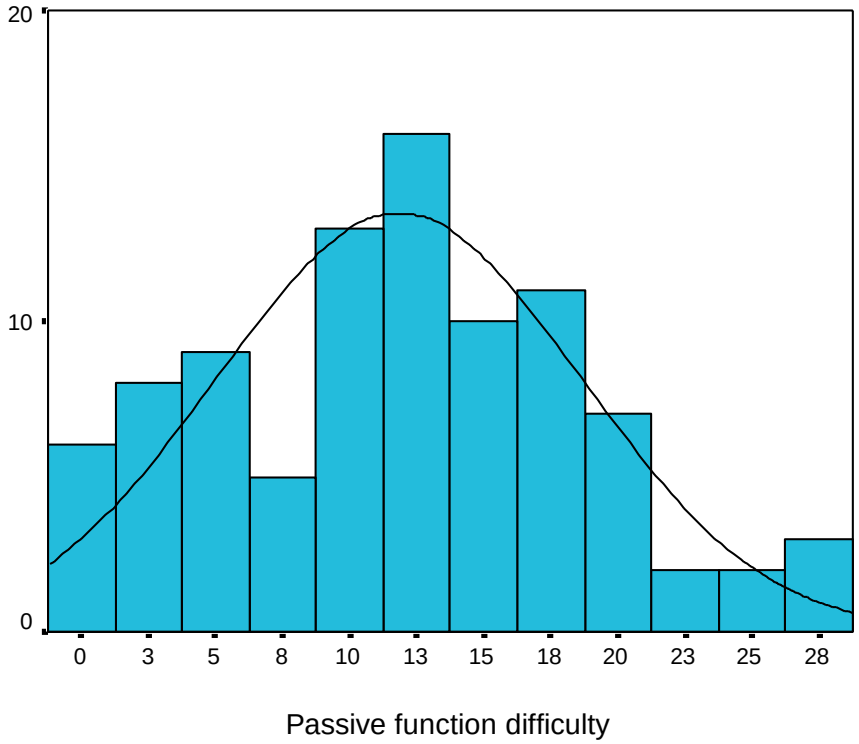
Measure		Time 1 Baseline	Time 2 1 day	Time 3 8 weeks	Time 4 16 weeks
G 1+2 (n=92)	**ArmA Passive**	12 (6.2-17)	13 (6-17)	-	-
	ArmA Active	48.5(35.5-52)	47.5(30-52)	-	-
	Barthel Index	12 (2-15)	-	-	-
G1 Only (n=58)	**ArmA Passive**	14 (10-18)	-	13 (7-17)	11 (7-15)
	ArmA Active	52 (48.7-52)	-	51 (48-52)	51 (48-52)
	Barthel Index	5 (0-15)	-	5 (0-15)	7 (0-15.5)
	LASIS Passive items	8 (4.3-15.6)	-	8 (2-11.5)	5.7 (1.5-14)
	LASIS Active items	0 (0-2.2)	-	0 (0-2.0)	0 (0-2.0)
	DASH Active items	105 (100-105)	-	105 (99-105)	105 (94.7-105)
	Modified Ashworth	4 (3-6)	-	2 (1-3)	2 (2-4)

Score ranges: ArmA Passive 0-28; Active 0-52, Barthel 0-20, LASIS Passive 0-36; Active 0-12, DASH Active 0-145, Modified Ashworth 0-5 (additional point +1).

7.4.2 Ceiling and floor effects

No ceiling or floor effects were identified in the Time 1 ratings on the passive function sub-scale. Scores were distributed over the range of the measure from 0 to 28, with an increased frequency in the centre of the scale. The modal score was 13, rated by 10 (11%) participants as shown in Figure 7.2.

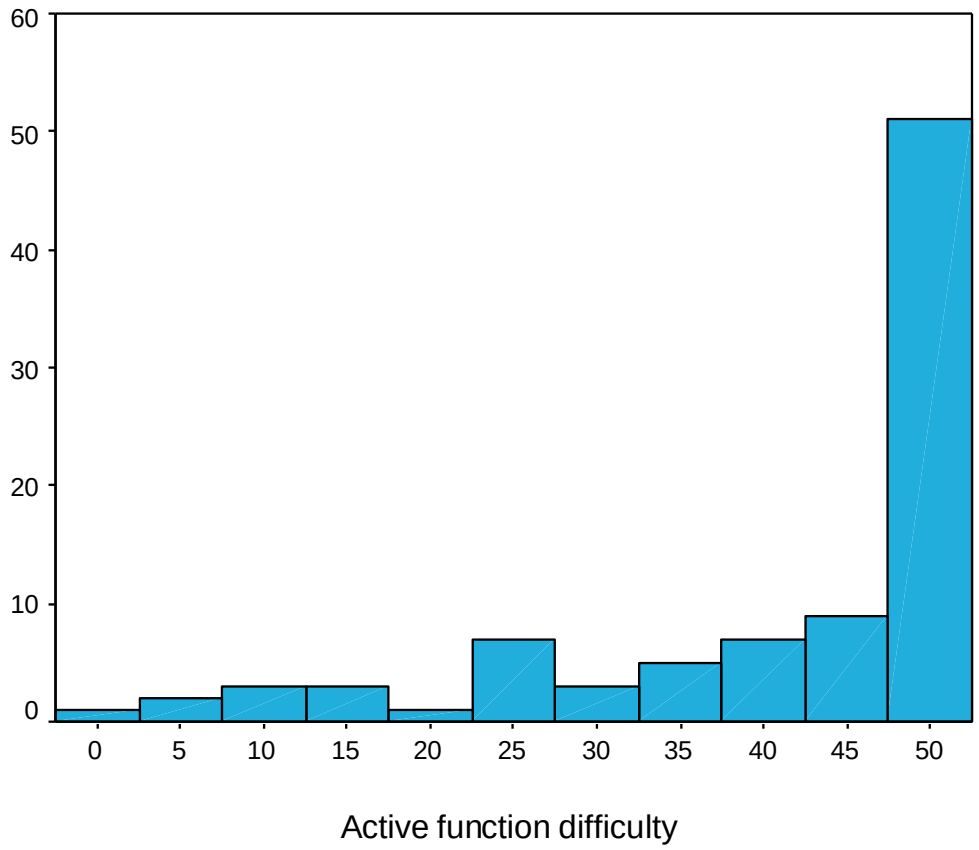
Figure 7.2 ArmA passive function score distribution across the scale



N = 92; Median = 12; Inter-quartile range = 6.2-17; Kolmogorov-Smirnov Z = 0.619; P = 0.84

In the active function sub-scale, a complete range of scores was produced at Time 1 from 0 to 52. However a ceiling effect occurred with 37% of scores for active function at the maximum point on the scale (52) shown in Figure 7.3, according to the criteria by Terwee and colleagues (Terwee et al. 2007).

Figure 7.3 ArmA active function score distribution across the scale



N = 92; Median = 48.5; Inter-quartile range = 35.5-52; Kolmogorov-Smirnov Z = 2.202; P = 0.003

The ceiling effect in the active function sub-scale became apparent during the initial phase of recruitment for Group 1 as 34 participants initially recruited were reporting maximum difficulty. This ceiling effect although reduced with the addition of Group 2, as expected was not entirely eradicated.

7.4.3 Construct validity

Construct validity was evaluated at Time 1 (baseline) by correlation of passive and active function sub-scales of the ArmA with the LASIS and DASH passive and active function items shown in Table 7.9.

Table 7.9 Correlation matrix between baseline ArmA (n=58), LASIS (n=57) and DASH (n=58).

Measure	ArmA Passive	ArmA Active
LASIS Passive items (n=57)	Rho = 0.5*	Rho = 0.23
LASIS Active items (n=57)	Rho = 0.02	Rho = 0.48*
DASH Active items (Items 1-21) (n=58)	Rho =-0.01	Rho = 0.63*

Key: All correlations used Spearman rank order correlation coefficient, *Significance $P < 0.01$

Convergent validity was shown by the significant correlation between the ArmA passive function sub-scale and the LASIS passive function items. In the active function sub-scale convergent validity was again shown by significant correlation between the ArmA active function items with the active function items from LASIS and DASH.

Divergent validity was shown by non-significant correlations of the ArmA passive function sub-scale with DASH and LASIS active items. The ArmA active function sub-scale also showed a non-significant correlation with LASIS passive function items.

In summary, the passive function sub-scale of the ArmA was found to correlate with passive function items in other measures and not to correlate with active function items. The active function sub-scale correlated with active function items such as DASH active items, but did not correlate with the passive function items. Construct validity is supported as per the recommendations of Terwee and colleagues (Terwee et al. 2007).

7.4.4 Unidimensionality and scaling

Principal component analysis

The passive function sub-scale had one component with an Eigenvalue above 1 using Kaiser's criteria (Kaiser 1960), which accounted for 54% of the variance while the second component accounted for only 12% of the variance shown in Table 7.10.

Table 7.10 Passive function; variance explained following principal component analysis (n=92)

Component	Initial Eigenvalues			
	Total	% of variance	Cumulative %	Random Eigenvalues*
1	3.79	54	54	1.41
2	0.88	12	67	1.22
3	0.68	10	76	1.08
4	0.63	9	85	0.97
5	0.46	6	92	0.88
6	0.29	4	96	0.78
7	0.27	4	100	0.64

*Monte Carlo analysis according to Horn's method.

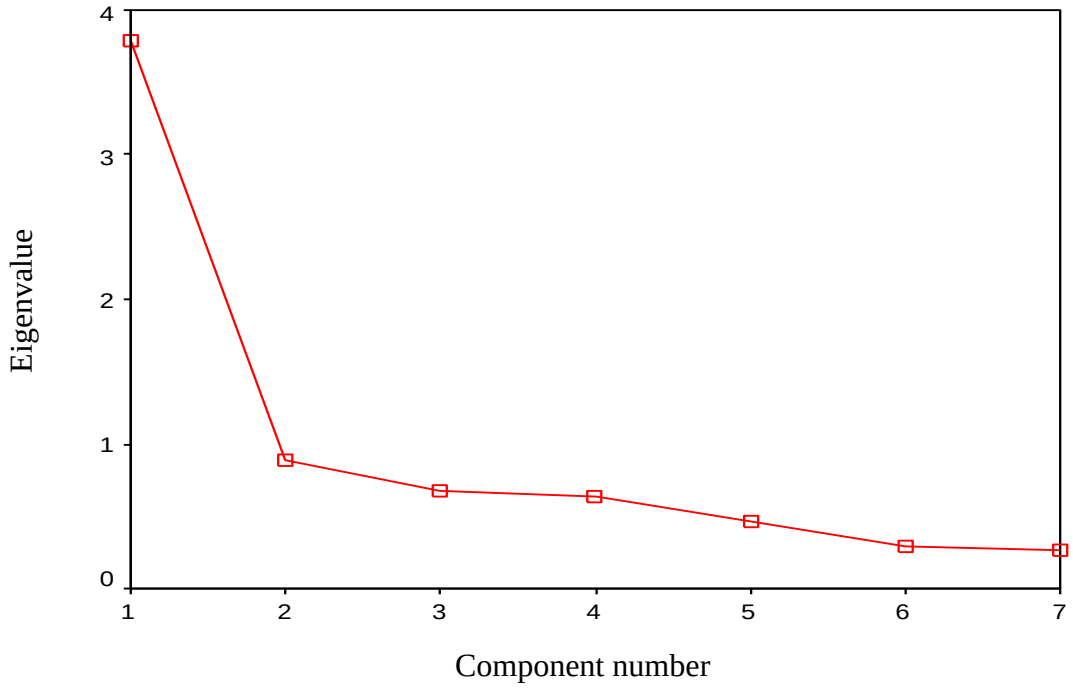
These results indicate that the first component has a much larger influence and suggests a relatively coherent single construct, which has been interpreted as summarising passive function. Table 7.11 displays loading of the individual measurement items onto the single principal component with an Eigenvalue above 1.

Table 7.11 Passive function item loadings onto the principal component

Item	Loading
1. Cleaning palm	0.876
2. Cutting finger nails	0.716
3. Putting on a glove	0.642
4. Cleaning armpit	0.801
5. Putting arm through sleeve	0.817
6. Putting on a splint	0.636
7. Position arm in sitting	0.622

Figure 7.4 shows a Scree plot of the principal components for the passive function sub-scale.

Figure 7.4 Scree plot of principal components - passive function sub-scale (n=92)



The Scree plot supports a single component interpreted as passive function. In addition a Monte Carlo analysis according to Horn's method of parallel analysis was undertaken and confirmed a single principal component (Horn 1965) (see Table 7.10).

The active function sub-scale had two components with an Eigenvalue above 1 using Kaiser's criteria (Kaiser 1960), which in part reflected the narrower spread of scores within this sample. However, the first component accounted for 71% of the variance while the second accounted for only 8% and was only just above 1. On conducting a Monte Carlo analysis according to Horn's method, only the first factor was retained indicating a single principal component (shown in Table 7.12).

Table 7.12 Active function; analysis of variance following principal component analysis (n=92)

Component	Initial Eigenvalues			
	Total	**% of variance**	**Cumulative %**	**Random Eigenvalues***
1	9.26	71	71	1.68
2	1.01	8	79	1.50
3	0.56	4	83	1.36
4	0.48	4	87	1.25
5	0.41	3	90	1.14
6	0.35	3	93	1.06
7	0.25	2	95	0.97
8	0.17	1	96	0.88
9	0.14	1	97	0.79
10	0.12	0.9	98	0.72
11	0.10	0.74	99	0.63
12	0.08	0.64	99	0.55
13	0.05	0.41	100	0.46

*Monte Carlo analysis according to Horn's method.

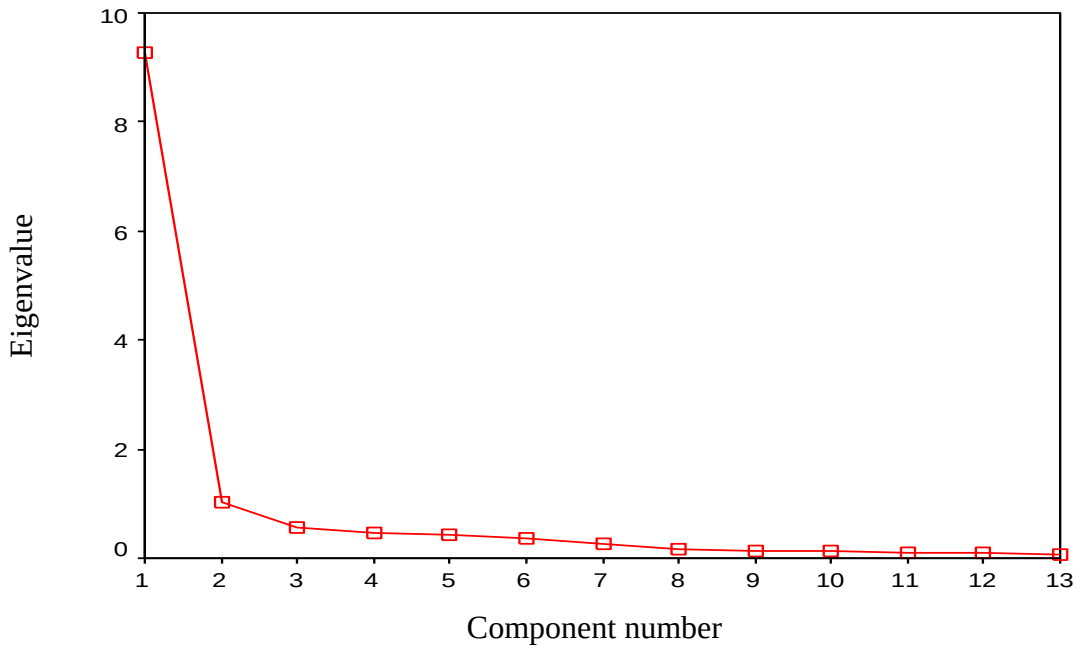
These results suggest that the first component corresponds with the construct of active function. Table 7.13 displays loading of the individual measurement items onto the two principal components with Eigenvalues above 1.

Table 7.13 Active function item loadings onto first and second principal components

Item	Loading 1st component	Loading 2nd component
1. Buttons on clothing	0.864	0.047
2. Pick up glass/bottle	0.852	0.125
3. Use a key in lock	0.889	-0.080
4. Write on paper	0.832	-0.316
5. Open a jar	0.858	0.059
6. Eat with knife & fork	0.856	-0.020
7. Hold object & use other hand	0.724	0.497
8. Effect of arm on balance when walking	0.523	0.726
9. Dial number on phone	0.930	-0.106
10. Tuck in shirt	0.935	-0.021
11. Comb hair	0.881	-0.141
12. Brush teeth	0.870	-0.208
13. Drink from cup	0.875	-0.185

The individual items load strongly onto the first component and generally weakly onto the second component. The item loadings therefore support the interpretation of the first component representing active function. Examination of the Scree plot (see Figure 7.5) for these data also further support a single principal component interpreted as active function.

Figure 7.5 Scree plot of principal components - active function sub-scale (n=92)



To further support single principal components for the active and passive function sub-scales, analysis was undertaken on the combined items from both sub-scales. In addition, Promax rotation was also performed with Kaiser normalisation to support the interpretation of a single principal component for each sub-scale.

Examination of the Eigenvalues for both sub-scales combined produces three Eigenvalues above 1 as per Kaiser's criteria shown in Table 7.14.

Table 7.14 Analysis of variance following principal component analysis for active and passive sub-scales combined (n=92)

Component	Initial Eigenvalues			
	Total	**% of variance**	**Cumulative %**	**Random Eigenvalues***
1	10.75	54	54	1.95
2	2.44	12	66	1.74
3	1.21	6	72	1.60
4	0.80	4	76	1.50
5	0.72	4	80	1.40
6	0.65	3	83	1.29
7	0.51	3	86	1.21
8	0.47	2	88	1.11
9	0.44	2	90	1.04
10	0.40	2	92	0.97
11	0.31	1	93	0.90
12	0.28	1	94	0.83
13	0.23	1	95	0.77
14	0.20	1	96	0.71
15	0.16	1	97	0.65
16	0.12	1	98	0.59
17	0.10	1	99	0.53
18	0.08	0.4	99	0.47
19	0.07	0.3	99	0.40
20	0.05	0.3	100	0.34

*Monte Carlo analysis according to Horn's method.

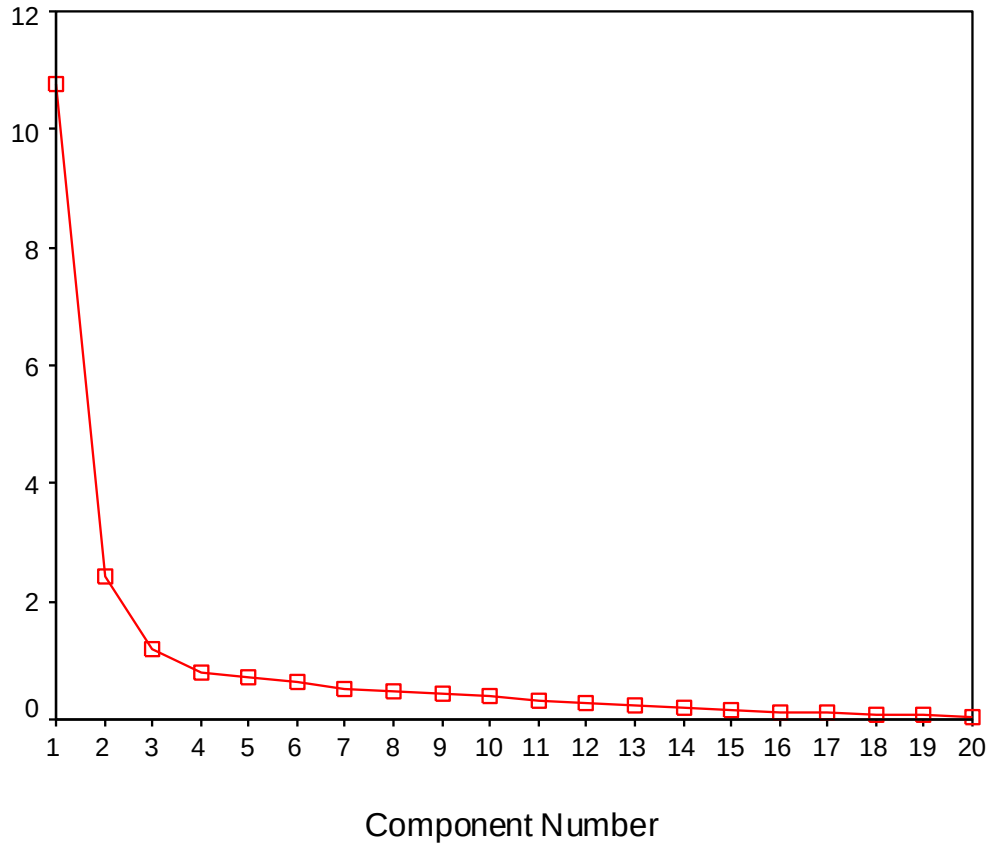
Three components had Eigenvalues above 1, however a Monte Carlo analysis was again conducted and confirmed two principal components when comparing the random values with those produced from the data. Table 7.15 displays the loading of the individual items onto the first and second components.

Table 7.15 Item loadings onto first and second principal components for combined active and passive function sub-scales

Item	Loading 1st component	Loading 2nd component
1. Cleaning palm	0.607	0.609
2. Cutting finger nails	0.494	0.469
3. Putting on a glove	0.542	0.327
4. Cleaning armpit	0.545	0.563
5. Putting arm through sleeve	0.522	0.651
6. Putting on a splint	0.461	0.463
7. Position arm in sitting	0.454	0.468
8. Buttons on clothing	0.848	-0.154
9. Pick up glass/bottle	0.847	-0.108
10. Use a key in lock	0.869	-0.187
11. Write on paper	0.795	-0.289
12. Open a jar	0.850	-0.116
13. Eat with knife & fork	0.829	-0.216
14. Hold object & use other hand	0.718	-0.047
15. Effect of arm on balance when walking	0.559	-0.215
16. Dial number on phone	0.905	-0.233
17. Tuck in shirt	0.919	-0.176
18. Comb hair	0.861	-0.204
19. Brush teeth	0.832	-0.302
20. Drink from cup	0.837	-0.296

Figure 7.6 shows the Scree plot of the principal components for the combined passive and active function sub-scales.

Figure 7.6 Scree plot of principal components for active and passive function sub-scales combined (n=92)



Promax rotation was performed on the combined scales and shown in Table 7.16.

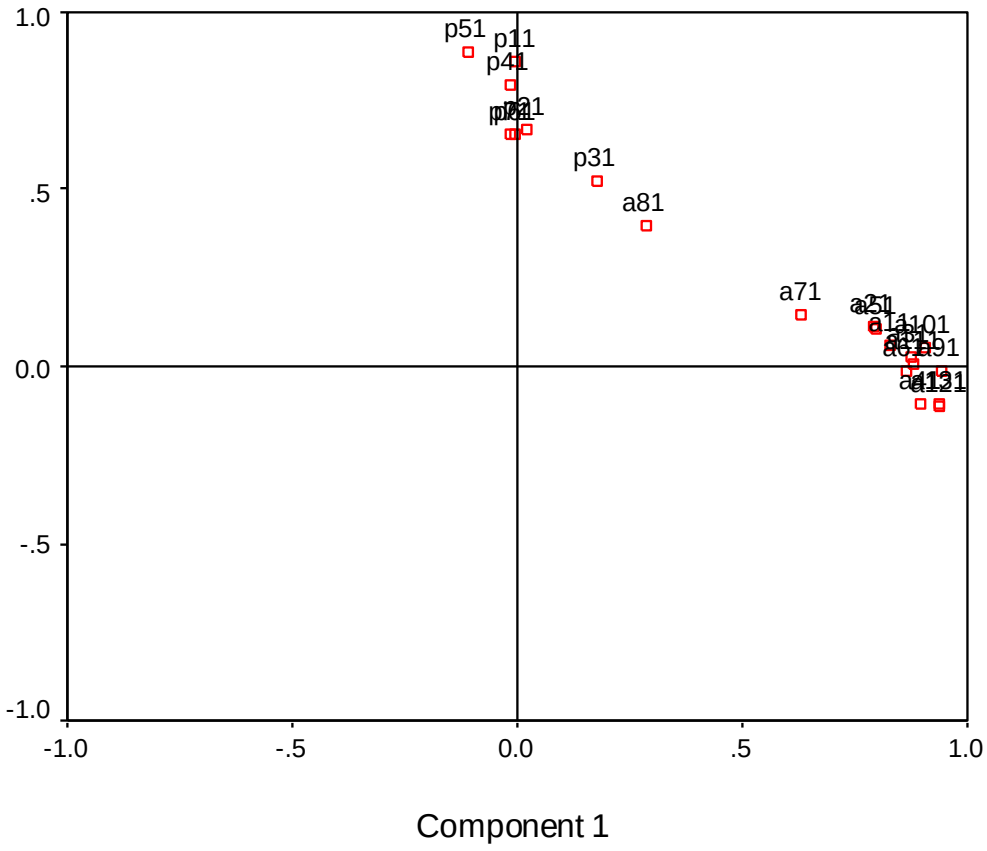
Table 7.16 Item loadings for combined active and passive function sub-scales following Promax rotation with Kaiser normalisation.

Item	Loading 1st component	Loading 2nd component
1. Cleaning palm	0.445	**0.860**
2. Cutting finger nails	0.368	**0.681**
3. Putting on a glove	0.449	**0.615**
4. Cleaning armpit	0.396	**0.784**
5. Putting arm through sleeve	0.353	**0.830**
6. Putting on a splint	0.338	**0.653**
7. Position arm in sitting	0.329	**0.651**
8. Buttons on clothing	**0.860**	0.493
9. Pick up glass/bottle	**0.848**	0.524
10. Use a key in lock	**0.889**	0.484
11. Write on paper	**0.841**	0.360
12. Open a jar	**0.853**	0.521
13. Eat with knife & fork	**0.856**	0.435
14. Hold object & use other hand	**0.708**	0.476
15. Effect of arm on balance when walking	0.492	**0.548**
16. Dial number on phone	**0.934**	0.477
17. Tuck in shirt	**0.934**	0.527
18. Comb hair	**0.884**	0.466
19. Brush teeth	**0.880**	0.377
20. Drink from cup	**0.883**	0.385

Bold indicates the highest loading of each item onto one of the two components

The results following Promax rotation were then plotted in two-dimensional space shown in Figure 7.7.

Figure 7.7 Component plot in two-dimensional space for active and passive function sub-scales combined (n=92)



The results following Promax rotation support two principal components with items in two groups on the plot in two-dimensional space. However, one item (a81) "difficulty with balance when walking due to your arm", is more closely related to passive function rather than active function. Principal component analysis suggested unidimensional sub-scales for passive and active function although this is a preliminary evaluation requiring further confirmation.

Mokken Analysis

The items in the passive function sub-scale produced an overall H coefficient for the scale of 0.48 with no individual item H-coefficients below 0.3 shown in Table 7.17.

Table 7.17 Mokken Analysis – passive function sub-scale (n=92)

Item	Summary per item	
	Mean	**Item H**
1. Cleaning palm	1.3	0.60
2. Cutting finger nails	2.2	0.48
3. Putting on a glove	1.6	0.53
4. Cleaning armpit	1.2	0.40
5. Putting arm through sleeve	1.6	0.55
6. Putting on a splint	2.2	0.43
7. Position arm in sitting	1.5	0.42
	Scale H	0.48
	Rho	0.85

van Schuur identifies that no item in a unidimensional scale should have an item H below 0.3, which is satisfied in the passive function sub-scale (van Schuur 2003). The overall H coefficient for the passive function total scale of 0.48 fits van Schuur's criteria for a moderately strong unidimensional scale (Molenaar et al. 2000; van Schuur 2003). Only one item, item 6, had a *crit* value above 40 (actual value 43) outside the recommended range of crit values (Molenaar et al. 2000) indicating possible violations of monotonicity. Although in this evaluation the aim was not to confirm double monotonicity, removal of items was explored. Removal of items 7 and 6 resulted in a 5 item scale with an overall item H of 0.56 indicating a strong scale and highest *crit* value of 15 indicating no violations of double monotonicity.

The items in the active function sub-scale yielded an overall H coefficient for the scale of 0.71 shown in Table 7.18.

Table 7.18 Mokken Analysis - active function sub-scale (n=92)

Item	Summary per item	
	Mean	**Item H**
1. Buttons on clothing	2.5	0.51
2. Pick up glass/bottle	3.1	0.64
3. Use a key in lock	3.4	0.73
4. Write on paper	3.1	0.72
5. Open a jar	3.3	0.73
6. Eat with knife & fork	3.2	0.72
7. Hold object & use other hand	3.2	0.79
8. Effect of arm on balance when walking	3.4	0.69
9. Dial number on phone	3.3	0.72
10. Tuck in shirt	3.3	0.80
11. Comb hair	3.3	0.73
12. Brush teeth	3.2	0.74
13. Drink from cup	3.4	0.76
	Scale H	0.71
	Rho	0.97

The overall H coefficient for the active function total scale of 0.71 fits van Schuur's criteria for a strong unidimensional scale (Molenaar et al. 2000; van Schuur 2003). The item with the third lowest H index was again "difficulty with balance when walking due to your arm". No item, had a *crit* value above 40 outside the recommended range of crit values (Molenaar et al. 2000) indicating no violations of monotonicity although this is likely, at least in part, to be due to the ceiling effect for this group.

Following Mokken analysis the unidimensionality of the two sub-scales is supported, although should be seen as preliminary evaluation due to the limited number of participants and the ceiling effect present in the active function data. However, the passive function sub-scale appears to be a unidimensional scale satisfying the monotone homogeneity model in its current form and may conform to double monotonicity (possibly with removal of items 6 and 7) when further evaluation is undertaken.

7.4.5 Internal consistency

Cronbach's alpha for passive function at Time 1 (baseline) was 0.85. This is within the range of 0.7 to 0.9 proposed by Nunnally and Bernstein (Nunnally and Bernstein 1994). The result also conforms to the quality criteria proposed by Terwee and colleagues (Terwee et al. 2007). Table 7.19 shows that internal consistency also remains high with the sequential removal of each item.

Table 7.19 Internal consistency – passive function (n=78)

Item Removed	Alpha	Median	Range	I.Q. Range
1. Cleaning palm	0.80	2	0-4	1-2
2. Cutting finger nails	0.83	2	0-4	1-4
3. Putting on a glove	0.84	3	0-4	2-4
4. Cleaning armpit	0.84	2	0-4	1-3
5. Putting arm through sleeve	0.81	2	0-4	1-3
6. Putting on a splint	0.84	2	0-4	0.75-3
7. Position arm in sitting	0.81	2	0-4	0-2.25
Total (all items included)	0.85			

Cronbach's alpha for active function at Time 1 (baseline) was 0.96 and reduced to 0.95 following the removal of two items shown in Table 7.20.

Table 7.20 Internal consistency – active function (n=78)

Item Removed	Alpha	Median	Range	I.Q. Range
1. Buttons on clothing	0.97	2	0-4	0-3
2. Pick up glass/bottle	0.95	4	0-4	1-4
3. Use a key in lock	0.95	3	0-4	1.75-4
4. Write on paper	0.96	3	0-4	1-4
5. Open a jar	0.96	3.5	0-4	2-4
6. Eat with knife & fork	0.96	3	0-4	1-4
7. Hold object & use other hand	0.96	2.5	0-4	1-4
8. Effect of arm on balance when walking	0.96	3.5	0-4	1.75-4
9. Dial number on phone	0.96	3	0-4	1.75-4
10. Tuck in shirt	0.96	3	0-4	2-4
11. Comb hair	0.96	3	0-4	2-4
12. Brush teeth	0.96	2	0-4	2-4
13. Drink from cup	0.96	4	0-4	2-4
Total (all items included)	0.96			

The active function sub-scale result indicates high internal consistency but is also indicative of item redundancy according to Terwee and colleagues (Terwee et al. 2007). However, the more plausible explanation for high internal consistency is the fact that many participants rated active function difficulty at maximum, thus producing the ceiling effect. Despite the ceiling effect, a full range of scores for all items in both passive and active function sub-scales was found in the data (see Tables 7.19 and 7.20). Medians were in the middle of the scale for the passive function items, but with a tendency towards the higher end of the scale for the active function items indicating more difficulty for many subjects.

In summary, Cronbach's alpha for the passive function and active function subscales showed high internal consistency. Mokken analysis also produces a version of internal consistency in the form of the overall Rho for the sub-scales (van Schuur 2003). The Rho for both sub-scales was greater than 0.80 supporting internal consistency using this method in addition to Cronbach's alpha (see Tables 7.17 and 7.18).

7.4.6 Test re-test reliability and agreement

Quadratic Weighted Kappa coefficients for the passive function scale were between 0.71 and 0.90. Percentage agreement ranged between 91.99 and 97.52; both Kappa coefficients and percentage agreement are shown in Table 7.21.

Table 7.21 Test re-test reliability (Time 1 to Time 2) passive function (n=78).

Item	% Agreement N=78	Kappa N=78	Standard error	Landis and Koch interpretation
1. Cleaning palm	96.23	0.82	0.1119	Almost perfect
2. Cutting finger nails	93.83	0.74	0.1130	Substantial
3. Putting on a glove	96.87	0.71	0.1132	Substantial
4. Cleaning armpit	91.99	0.80	0.1107	Almost perfect
5. Putting arm through sleeve	96.31	0.75	0.1123	Substantial
6. Putting on a splint	97.52	0.90	0.1132	Almost perfect
7. Position arm in sitting	96.47	0.86	0.1125	Almost perfect

The Kappa coefficient for passive function conformed to "substantial" or "almost perfect" criteria for all items according to Landis and Koch (Landis and Koch 1977).

Quadratic Weighted Kappa coefficients for the active function scale were between 0.70 and 0.94. Percentage agreement ranged between 92.15 and 98.72; both Kappa coefficients and percentage agreement are shown in Table 7.22.

Table 7.22 Test re-test reliability Time 1 to Time 2 active function (n=78).

Item	% Agreement N=78	Kappa N=78	Standard Error	Landis and Koch interpretation
1. Buttons on clothing	96.23	0.94	0.1126	Almost perfect
2. Pick up glass/bottle	96.23	0.89	0.1130	Almost perfect
3. Use a key in lock	96.79	0.76	0.1130	Substantial
4. Write on paper	95.11	0.79	0.1130	Substantial
5. Open a jar	94.87	0.81	0.1132	Almost perfect
6. Eat with knife & fork	94.79	0.81	0.1131	Almost perfect
7. Hold object & use other hand	97.52	0.87	0.1131	Almost perfect
8. Effect of arm on balance when walking	98.72	0.86	0.1128	Almost perfect
9. Dial number on phone	97.20	0.78	0.1132	Substantial
10. Tuck in shirt	97.28	0.88	0.1131	Almost perfect
11. Comb hair	92.15	0.86	0.1130	Almost perfect
12. Brush teeth	96.39	0.70	0.1127	Substantial
13. Drink from cup	96.55	0.85	0.1118	Almost perfect

Weighted Kappa coefficients for the active function sub-scale also conformed to "substantial" or "almost perfect" agreement criteria according to Landis and Koch (Landis and Koch 1977) (see Table 7.22). Again, the tendency towards ceiling effect in

the active function sub-scale has probably over emphasised the degree of agreement. Using the criteria of Landis and Koch both sub-scales are given a positive rating for reliability according to Terwee and colleagues (Terwee et al. 2007).

7.4.7 Responsiveness

Primary classification of response following intervention (BTX and PT) at Time 3 (8 weeks) was determined by CCR. As further evidence of responsiveness and longitudinal validity, response at Time 4 (16 weeks) is also presented. The categorisation of response identified by CCR and corroborated by individual goal attainment recorded by GAS (goals achieved as predicted or over achieved), is presented in Table 7.23.

Table 7.23 Response identified at 8 and 16 weeks by CCR and GAS.

Measure	8 Weeks	16 Weeks
CCR	(n=52)	(n=48)
Responders	41	36
Non-responders	11	12
Unknown	6	10
GAS	(n=53)	(n=47)
Responders	42	36
Non-responders	11	11
Unknown	5	11

The number of responders identified by CCR and GAS was similar with 41 and 42 respectively at Time 3 (8 weeks). The same participants were identified by both CCR and GAS, apart from one participant who did not have CCR recorded by his treating therapists. At Time 4 (16 weeks) both CCR and GAS identified the same 36 participants as responders.

The responsiveness of the ArmA passive and active function sub-scales was evaluated by comparing responders with non-responders for passive and active sub-scales at Time 3 (8 weeks) using the Mann-Whitney U test. To further support these findings and provide evidence of longitudinal validity, the evaluation was repeated at Time 4 (16 weeks). To allow comparison of the ArmA with the other measures, responsiveness of

LASIS passive items, LASIS active items, Barthel index and DASH active items, were also evaluated. The results of these analyses are shown in Table 7.24.

Table 7.24 Responder Vs non-responder change in the ArmA, LASIS, Barthel and DASH (n=51).

Measure	Responders Vs non-responders (8 weeks)	Responders Vs non-responders (16 weeks)
ArmA passive	**U = 98.5; p = 0.01**	**U = 29.5; p = 0.003**
ArmA active	U = 163.5; p = 0.35	U = 77.5; p = 0.21
LASIS passive items	U = 127.0; p = 0.07	U = 263.5; p = 0.20
LASIS active items	U = 167.0; p = 0.39	U = 210; p = 0.17
DASH active items	U = 176.5; p = 0.92	U = 136.0; p = 0.47
Barthel Index	U = 200.5; p = 0.17	U = 117.5; p = 0.06

Significant results in bold

The ArmA identified a significant difference between responder and non-responder groups for the passive function sub-scale at Time 3 (8 weeks) for primary evaluation of responsiveness and at Time 4 (16 weeks). A significant difference was not shown for the active function sub-scale. The other measures did not show a significant difference between responders and non-responders at Time 3 (8 weeks) or Time 4 (16 weeks). For graphical representation of individual patient level change in the ArmA passive function sub-scale at baseline, 8 weeks and 16 weeks see Appendix 19.

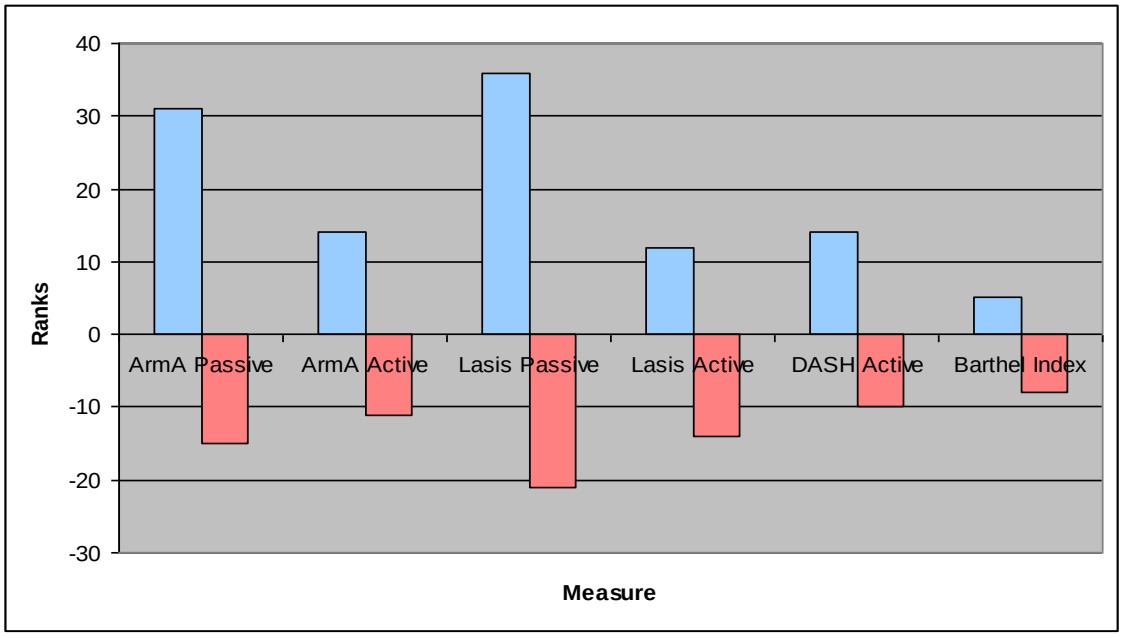
Cohen's effect size and standard response mean were also calculated for ArmA, LASIS, Barthel Index and DASH active items shown in Table 7.25. As already discussed in chapter 3, the ordinality of the data brings into question the use of these methods with the ArmA data. However, these methods have been used here to allow comparison between the different measures applied in the evaluation, to inform discussion of findings and the relationship of the ArmA to other measures.

Table 7.25 Effect size and standard response mean of the ArmA sub-scales, LASIS active and passive items, Barthel Index and DASH active items (n=51).

Measure	Effect Size	Standard Response Mean
ArmA passive	0.29	0.30
ArmA active	0.21	0.16
LASIS passive	0.23	0.20
LASIS active	0.04	0.03
DASH active	0.07	0.08
Barthel	0.07	0.12

Cohen's effect size and standard response mean were small for the LASIS active items, Barthel Index, and the DASH active items. Although somewhat greater for the ArmA active and passive function sub-scales and LASIS passive items, they were still low. Given the non-parametric nature of the data, the number of positive and negative rank differences are also presented for each measure from baseline to 8 weeks (see Figure 7.8).

Figure 7.8 Positive and negative rank differences from baseline to 8 weeks



The distribution of data was assessed using the Kolmogorov-Smirnov test (also see section 7.4.2, Ceiling and floor effects; pages 208-210) and was normal for ArmA passive function (Z = 0.619; P = 0.84) and LASIS passive items (Z = 1.192; P = 0.12).

Distribution was not normal for ArmA active function (Z = 2.202; P = 0.003), LASIS active function items (Z = 2.300; P = 0.003), DASH active function items (Z = 2.498; P = 0.003) or Barthel index (Z = 1.471; P = 0.026). The results again support greater, yet still small change in the ArmA passive function sub-scale and the LASIS passive function items. However given the limitations of the sample for active function and the non-parametric nature of the data, conclusions about the responsiveness are preliminary.

7.4.8 Interpretability

Minimal Important Change (MIC) was calculated using: a criterion-based method and a distribution-based method (see Interpretability, page 203). The results of calculating MIC using both these methods for the passive and active sub-scales are presented in Table 7.26.

Table 7.26 Minimal Important Change calculated using criterion and distribution methods (n=51).

Sub-scale	Criterion-based method	Distribution-based method
ArmA passive	2.5	3.0
ArmA active	1.1	2.5

Given that no significant difference was identified between the responder and non-responder group for active function the calculated MIC has limited meaning and will require further evaluation. The ArmA passive function MIC will also need further evaluation once fundamental measurement properties have been demonstrated to allow confidence in the results of this parametric approach.

To evaluate these preliminary predictions of MIC, sensitivity and specificity was calculated based on grouping of responders by CCR at 8 weeks for ArmA passive function (see Interpretability, page 203). Calculation was undertaken for a 1-point change in ArmA, a 2-point change in ArmA and a 3-point change in ArmA. The results are presented in Table 7.27.

Table 7.27 Sensitivity and Specificity of the ArmA according to classification of CCR at 8 weeks.

Change from Baseline	1 point change	2 point change	3 point change
Sensitivity	0.69	0.55	0.45
Specificity	0.58	0.91	0.91
Positive predictive value	0.63	0.45	0.37
Negative predictive value	0.37	0.55	0.63

These results indicate optimal sensitivity was achieved with a 1-point change in ArmA, but that specificity was improved with a 2-point change in ArmA.

7.4.9 Feasibility

Ease of completion was rated as Very easy, Easy or Moderate (on a scale from Very easy to Very difficult) by 90% of patients or carers, shown in Table 7.28.

Table 7.28 Ratings for ease of the ArmA completion (n=56)

Ease of Completion	Number (Percentage)
Very easy	15 (26 %)
Easy	26 (45 %)
Moderate	11 (19 %)
Difficult	3 (5 %)
Very Difficult	1 (2 %)
Missing	2 (3 %)

The ArmA was completed in 83% (n=48) of respondents in 10 minutes or under, presented in Table 7.29.

Table 7.29 Ratings for time taken by patients and carers to complete the ArmA (n=56)

Time to complete	Number (Percentage)
Under 5 minutes	27 (47%)
5-10 minutes	21 (36%)
11-15 minutes	3 (5%)
16-20 minutes	3 (5%)
Over 20 minutes	2 (3%)
Missing	2 (3%)

Only two questionnaires took more than 20 minutes to complete (see Table 7.29). The small numbers of participants (n=10) taking more than 10 minutes to complete the ArmA may have been due to confusion between the ArmA and the whole questionnaire pack (including all the other measures) for this minority of subjects.

Relevance of the overall scale was rated by 77% of respondents as Very relevant to Moderately relevant. The active function sub-scale was rated as Very useful to Moderately useful by 71% of respondents and the passive function subscale by 88% of respondents shown in Table 7.30.

Table 7.30 Ratings of relevance or usefulness by patients and carers (n=56)

Relevance or Usefulness	Whole ArmA (Relevance)	Active Function (Usefulness)	Passive Function (Usefulness)
Very Relevant/ useful	15 (26%)	13 (22%)	16 (28%)
Relevant/ useful	17 (29%)	18 (31%)	19 (33%)
Moderate	13 (22%)	10 (17%)	15 (26%)
Little relevance/ usefulness	8 (14%)	11 (19%)	4 (7%)
No relevance / usefulness	3 (5%)	4 (7%)	2 (3%)
Missing	2 (3%)	2 (3%)	2 (3%)

In table 7.30 relevance or usefulness of the ArmA are given in the left hand column. Relevance is rated for the whole ArmA (passive and active function sub-scales); usefulness is rated for active function and passive function sub-scales individually. Although over a quarter of respondents (26%) said the active function sub-scale was of little or no relevance (due primarily to the dependency of the group), only 10% said this of the passive function sub-scale.

In summary, completion of ArmA was rated as Very easy to Moderately easy by 90 % of respondents and 83% of respondents completed the ArmA in 10 minutes or under. The passive and active function sub-scales of the ArmA are feasible to use for respondents as a self-completion questionnaire.

7.5 Results - Evaluation of functional change: A cohort study

7.5.1 Interventions applied

The cohort study involved Group 1 (G1) participants only. Injection of BTX was undertaken to muscles acting over the shoulder, elbow, wrist and hand. All participants received BTX intervention shown in Table 7.31.

Table 7.31 BTX intervention categorised by joint.

BTX intervention				
		Shoulder	**Elbow**	**Wrist/Hand**
Number of patients injected per region		9 (15%)	26 (45%)	29 (50%)
Muscles injected		Pectoralis Major Infraspinatus Latissimus Dorsi	Brachialis Brachioradialis Biceps Brachii Triceps Brachii	FDS FDP FCU FCR FPL
Dose range	**Dysport**	150-1000u	150-1000u	300-1000u
	Botox	-	50-200u	100-200u

FDS - Flexor Digitorum Superficialis; FDP -Flexor Digitorum Profundus; FCU - Flexor Carpi Ulnaris; FCR -Flexor Carpi Radialis; FPL - Flexor Policis Longus

In some patients, injections were applied to muscles acting over two joints. Dysport and Botox BTX-A products were applied, although no Botox injection was applied to muscles of the shoulder (see Appendix 18 for details of BTX dose for each participant).

All participants also received PT interventions shown in Table 7.32.

Table 7.32 Physical therapy interventions.

	Therapy intervention categories					
	Splinting	**Serial Casting**	**Positioning**	**Passive Stretch**	**FES**	**Task Practice**
Number patients	34 (59%)	11 (19%)	14 (24%)	34 (59%)	2 (3%)	7 (12%)
Specific intervention (Dose range)	Thermo-plastic (6-10 hrs Daily) Palm protector (6-24 hrs Daily) Non-custom e.g. Ultraflex (6-10 hrs Daily)	Non removable (2-4 Days) Removable (6-24 Hours)	Wheelchair tray (4-8 hrs Daily)	(1-3 X Daily 10 Reps. 30 Second holds)	(20 Mins Daily)	MCIMT (20 Mins X 3 Daily) Self exercise (2-4 X Daily) Therapy session exercise (4-10 X Weekly)
Region applied	Wrist/Hand Elbow	Elbow	Whole limb	Wrist/Hand Elbow Shoulder	Whole limb	Whole limb

Functional Electrical Stimulation (FES); Modified Constraint Induced Movement Therapy (MCIMT).

Injection of BTX was undertaken in conjunction with PT interventions, which were assigned to the six categories shown in Table 7.32. The most common PT interventions applied were splinting and passive stretch in 34 (59%) patients for both interventions. PT interventions were rarely provided in isolation, with most participants receiving at least two interventions in combination (mean number of interventions 1.7). In 17 (29%) participants single physical therapy interventions were provided, in these instances intervention was most commonly splinting to maintain range of movement (see Appendix 18 for details of PT dose for each participant).

7.5.2 Descriptive statistics

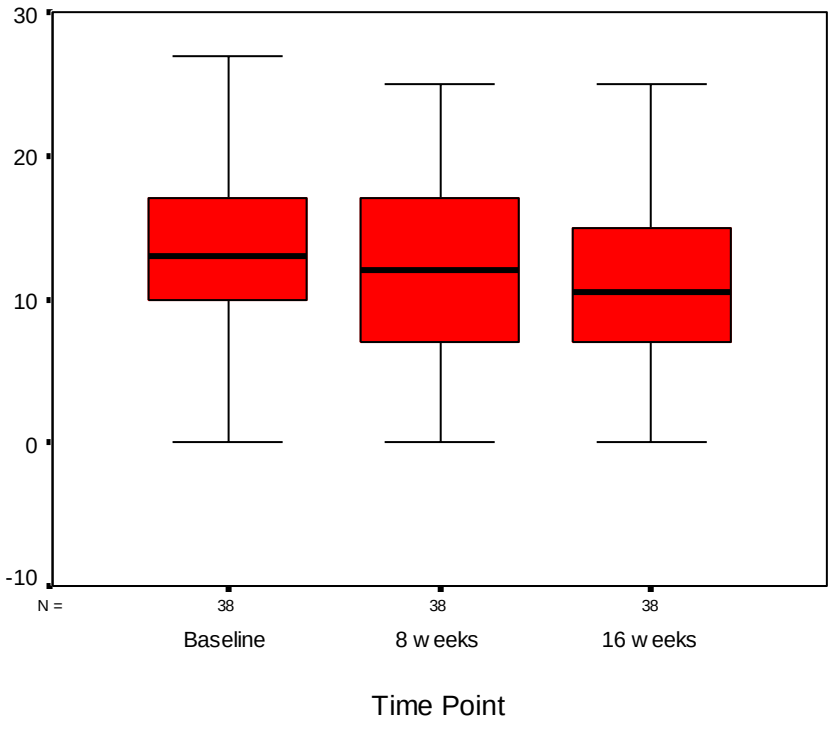
The clinical and demographic characteristics of participants were presented previously in Table 7.4 (page 206). Group 1 participants had a mean Barthel Index score of 7.0, (SD 7.7) indicating a high level of disability and dependence in ADL.

The ArmA, MAS, LASIS and DASH active ratings at Times 1, 3 and 4 are shown in Table 7.8 (page 208). A reduction in median ArmA passive function score from baseline to 8 and 16 weeks suggests reduced difficulty in performing passive function tasks (i.e. caring for the affected upper limb). The median MAS score also reduced from Time 1 (baseline) to Time 3 (8 weeks) and Time 4 (16 weeks) indicating a reduction in median spasticity. Change in LASIS passive function median score is also shown from Time 3 (8 weeks) to Time 4 (16 weeks), but no change in LASIS active function or DASH active function. A single point change is shown in ArmA active function median score from baseline to 8 weeks and maintained at 16 weeks.

7.5.3 Evaluation of functional change

ArmA passive function was compared between Time 1 (baseline), Time 3 (8 weeks) and Time 4 (16 weeks) and is displayed in Figure 7.9.

Figure 7.9 ArmA passive function change from baseline to 8 and 16 weeks (n=38)

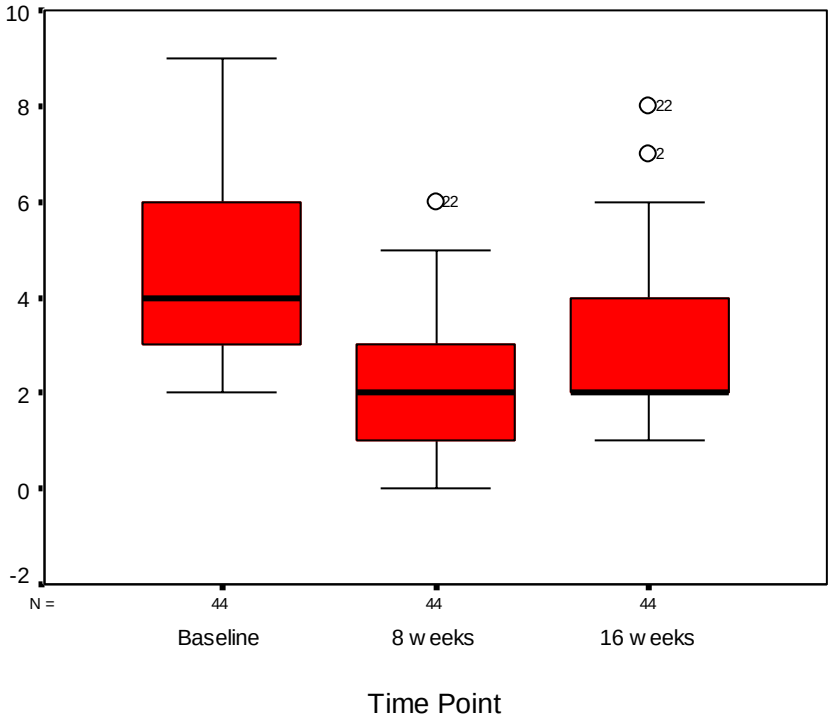


Box plots for passive function ArmA, following the Wilcoxon signed-ranks test, are shown across time points. Difficulty in passive function is reduced at 8 weeks and

appears to reduce further at 16 weeks. The range for all three-time points overlap due to variation within the group. Data are presented for 38 participants due to missing data at 16 weeks.

Composite Modified Ashworth was also compared between Time 1 (baseline), Time 3 (8 weeks) and Time 4 (16 weeks) and is displayed in Figure 7.10.

Figure 7.10 Composite Modified Ashworth change from baseline to 8 and 16 weeks (n=44)



Box plots for composite MAS, following the Wilcoxon signed-ranks test, are also shown at each time point. According to these measures, spasticity is reduced at 8 weeks and although still reduced increases again at 16 weeks. The ranges do not overlap between Time 1 (baseline) and Time 3 (8 weeks) but do overlap between Time 3 and Time 4. Data are presented for 44 participants due to missing data.

The Friedman test was applied to compare change between Time 1 (baseline), Time 3 (8 weeks) and Time 4 (16 weeks) for ArmA passive and active function, composite Modified Ashworth score measuring spasticity, LASIS passive and active function items and DASH active function items. The Wilcoxon test was then applied as per the

recommendations of Altman (Altman 1991)(p. 203-5) to compare Baseline to Time 3 (8 weeks) and Baseline to Time 4 (16 weeks) separately and shown with the results for the Friedman test in Table 7.33.

Table 7.33 Change in the ArmA, MAS, LASIS and DASH from baseline to 8 weeks (n=51) and baseline to 16 weeks (n=38).

Measure	Friedman test Baseline to 8 and 16 weeks	Wilcoxon Baseline to 8 weeks	Wilcoxon Baseline to 16 Weeks
ArmA passive	X^{r2} = 11.9***	Z = -2.1*	Z = - 2.6**
ArmA active	X^{r2} = 0.5	Z = -1.2	Z = -1.3
MAS	X^{r2} = 66.6***	Z = -6.0***	Z = -5.3***
LASIS Passive items	X^{r2} = 6.4*	Z = -1.7	Z = -1.5
LASIS Active items	X^{r2} = 2.0	Z = -0.1	Z = -1.0
DASH Active items	X^{r2} = 1.4	Z = -0.7	Z = -1.4

***Significant at P < 0.005; **Significant at P < 0.01; *Significant at P < 0.05

The results for the Friedman test show a significant decrease between Time 1 (baseline) and Time 3 and Time 4 (8 and 16 weeks) for the ArmA passive function sub-scale (X^{r2} 11.9; n = 51; p < 0.005). Due to the highly significant result from the Friedman test the risk of type one error was deemed small and the Wilcoxon test was applied. A significant difference was also identified for composite Modified Ashworth (X^{r2} 66.6; n = 51; p < 0.005) and in LASIS passive function items (X^{r2} = 6.4*; n = 51; p < 0.05). No significant difference was identified for the ArmA active function sub-scale (X^{r2} 0.5; n = 51; p > 0.05), LASIS active function items or DASH active function items.

A significant difference was identified between Time 1 (baseline) and ArmA passive function at Time 3 (8 weeks), with improvement in passive function (Z -2.1; n = 51; $p < 0.05$). Comparison of Time 1 (baseline) to Time 4 (16 weeks), showed that the difference from baseline was maintained (Z - 2.6; n = 52; $p < 0.01$), and did not show a statistically significant change from 8 weeks (Z -1.0; n = 51; $p > 0.05$). Significant difference was also identified for composite Modified Ashworth scores between Time 1 (baseline) and Time 3 (8 weeks) (Z -6.0; n = 51; $p < 0.005$), Time 1 (baseline) and Time 4 (16 weeks) (Z -5.3; n = 44; $p < 0.005$) and 8 weeks to 16 weeks (Z -3.6; n = 44; $p < 0.005$).

These results indicate that passive function is improved following BTX administration and PT intervention by 8 weeks and is maintained at 16 weeks despite the physiological effects of the BTX gradually reducing over this time period. The passive function sub-scale of the ArmA showed significant change at 8 weeks post intervention, which was maintained at 16 weeks following intervention. The null hypothesis is therefore rejected.

7.6 Discussion

Table 7.34 summarises the development (Chapters 4, 5, and 6) and evaluation of the ArmA (Chapter 7) in relation to the Quality Criteria produced by Terwee and colleagues and additional elements for evaluation of dimensionality, measurement scaling and feasibility.

Table 7.34 Summary of ArmA psychometric properties.

Attribute	Criteria	Evaluation
Validity	***The degree to which the instrument measures what it purports to measure***	
	Face	Confirmed during pilot testing with patients and carers before main psychometric evaluation and during item selection by inclusion of patient identified items.
	Content	***Aim; population and target concepts:*** The ArmA was designed to provide a low burden measure of difficulty in active and passive arm function for patients undergoing spasticity management in the upper limb. ***Item selection and reduction;*** Item selection used a systematic review and patient selected items. Item reduction was undertaken using a Delphi consensus process with specialist clinicians and confirmed with a wider consultation. Pilot testing was undertaken by patients and carers, feedback was provided on content and presentation. ***Interpretability of items:*** Items included in the questionnaire are short and simple and were developed in accordance with recommendations on language. Understanding was confirmed during pilot testing (Terwee et al. 2007).
	Criterion-related	Not testable - no accepted gold standard measure for comparison currently exists. Comparison using direct observation or activity analysis is impractical and raises problems of bias due to the presence of the observer.
	Construct (Concurrent)	***Convergent***: Passive function sub-scale correlated with passive function items from LASIS (Rho 0.5) and active function was correlated with active function items from LASIS (Rho 0.48) and active function items from DASH (Rho 0.63). ***Divergent:*** Passive function did not significantly correlate with DASH active items and LASIS active items. The active function subscale was not correlated with LASIS passive function items. Construct validity is supported as per the recommendations of Terwee and colleagues with all correlations supporting the relationship of ArmA as predicted to the comparison measures (Terwee et al. 2007).
	Unidimensionality	Principal component analysis indicated one main component in each of the subscales and Mokken analysis confirmed the unidimensionality of the sub-scales - H index for the passive function 0.48 (rho 0.85) and active function 0.71 (rho 0.97).
Scaling		Mokken Analysis using a monotone homogeneity model confirmed preliminary ordinal scaling properties in the passive function sub-scale. H index for the passive function was 0.48 indicating a medium scale (between 0.4 and 0.49) to differentiate persons.
Reproducibility	***The degree to which the instrument is free from random error***	
	Agreement	Percentage agreement ranged between 91.99 - 96.87 for the passive function scale and 92.15 - 97.52 for the active function scale.
	Test re-test Reliability	Quadratic Weighted Kappa coefficients for the passive function sub-scale were between 0.71 - 0.90 and 0.70 - 0.94 for the active function sub-scale, constituting a positive rating for reliability (above 0.70 for Quadratic-weighted Kappa (Terwee et al. 2007)).

Responsiveness	***Ability to detect change over time where actual changes occur***	
	Change: Baseline prior to intervention and follow-up 8 weeks	In this evaluation, a significant difference was identified between responders and non-responders using the ArmA passive function sub-scale at 8 weeks (U=98.5, P = 0.01), but not the active function sub-scale. SRM and ES were greater for ArmA than other measures but still low.
Interpretability	***The degree to which easily understood meaning can be assigned to the quantitative scores***	
	Clinical meaning	MIC was calculated using a criterion based method (2.5) and a distribution based method (3.0) for the passive function sub-scale. The higher of the two results (3.0) is accepted as the preliminary value of MIC, requiring further evaluation with interval scaling.
Floor/ceiling effects		No significant floor or ceiling effects in the passive function sub-scale, but >15% ceiling effect in the active function sub-scale in the patient group used in this evaluation.
Feasibility and Burden	***The time, effort or other demands of completing the measure***	
	Time to administer	The ArmA was completed in 10 minutes or under by 83% of respondents.
	Ease of use	Ease of completion was rated as very easy, easy or moderate by 90 % of patients or carers.
	Relevance	Relevance of the overall scale was rated by 77% of respondents as very relevant to moderately relevant. The active function sub-scale was rated as very useful to moderately useful by 71% of respondents and the passive function subscale by 88% of respondents.
	Acceptability	Completion of ArmA was rated as very easy to moderate by 90% of respondents and 77% of respondents rated it as relevant with a mean time for completion of between 5-10 minutes.
	Value	The passive function subscale was rated by 88% of respondents as useful (very useful, useful or moderately useful). The active function subscale was rated by 71% of respondents as useful (very useful, useful or moderately useful).
Alternative modes of administration		The ArmA has been administered, during testing, as a self-completion questionnaire or as an interview (face-to-face or over the telephone). Only a small minority completed by interview or telephone and further validation of these methods will be needed.
Cultural and language adaptations		None currently available

MIC - Minimal Important Change; SRM – Standard Response Mean; ES – effect Size

The ArmA provides a valid and reliable measure of difficulty in passive arm function for spasticity management intervention in the hemiparetic upper limb. However, the fundamental measurement properties of the ArmA need further evaluation for passive and active sub-scales. Further establishment of measurement properties with Rasch analysis will allow further subsequent analysis of properties such as MIC. The evaluation of the active function sub-scale has been significantly limited by the ceiling effect seen in the study group and requires further evaluation in a more able study population.

The psychometric evaluation is discussed in Section 7.6.1 and the evaluation of functional change (cohort study) is discussed in Section 7.6.2.

7.6.1 Evaluation of the psychometric methods

While key psychometric concepts have been addressed, some limitations to the methods became apparent through the evaluation process. Moreover, some properties would also benefit from further evaluation.

7.6.1.1 Construct validity

Construct validity is supported, with confirmation of predicted moderate correlations of the ArmA with comparison measures. The use of DASH as a comparison measure for construct validity can be criticised due its lack of validation in individuals with neurological impairment. However, its potential strength for measurement of active function is in its design for self-completion in common with the ArmA. It also contains a number of very relevant items for patients with hemiparesis. Conversely, DASH contains a number of much higher-level functional items, which were unlikely to show change in all but the most able participants with impairment of neurological origin. Other alternatives such as MAL-14, MAL-26 or MAL-28 are not currently designed for self-completion. Both MAL-14 and MAL-28 were pilot tested in standard format and an adapted form for self-completion, before construction of the measures pack. In both versions of MAL respondents are asked to score both ‘Quality of Movement’ and ‘amount of use’ for the affected upper limb as per the instructions for this measure. In the pilot testing, these concepts were difficult for patients and carers to understand, which was particularly apparent for ‘quality of movement’ which was often not

completed. The decision not to use the MAL in the ArmA validation was made on this basis.

In the evaluation by van der Lee and colleagues (2004) correlations were identified between the MAL-14 and the Action Research Arm Test (ARAT) at baseline, which was a similar approach to that taken for convergent validity with the ArmA. However, a weakness in the approach for the ArmA is the use of items, taken from measures, for comparison (e.g. passive function items from LASIS) rather than use of the whole measure. This method was necessary to compare relevant items with the sub-scales of the ArmA, since other measures exactly matching the ArmA sub-scales (i.e. 'gold-standard' measures) could not be found (see Chapter 4).

Uswatte and colleagues also examined construct validity for the MAL-14 using accelerometer recordings (Uswatte et al. 2005). Accelerometers were worn for three days at baseline and at the point of outcome evaluation. Pearson correlations were identified between accelerometer recordings and baseline MAL-14 and change in MAL-14 (Uswatte et al. 2005). This has parallels with the approach taken in the ArmA, in evaluating hypothesised change over time, but differences in the manner of measurement. However, in ArmA the statistical method of Spearman correlation was used because of the non-parametric nature of the data, which would seem a more appropriate method as per the criteria of Terwee and colleagues (2007) than Pearson correlation used for evaluation of MAL-14.

In the evaluation of MAL-28 by Uswatte and colleagues, extensive use was made of accelerometers placed on each wrist to record participants function (Uswatte et al. 2006). Correlations were identified between MAL-28 and accelerometer recordings of the affected upper limb. Convergence was also seen with MAL-28 and the Stroke Impact Scale – hand function section completed by both patients and carers. Divergence was identified between accelerometer recordings on the unaffected upper limb and the Stroke Impact Scale – mobility section completed by patients. However caregiver completed MAL-28 did correlate with accelerometer recordings for the unaffected hand. Use of accelerometer recording is possibly the next best method to direct observation of activity but is still time consuming, requires significant

commitment from the participant and can still only record information over a limited time period (72 hours for MAL-28). Further evaluation of the ArmA could include comparison with accelerometer recordings, which should give further evidence for construct validity, but may have limited value as construct validity is now well supported.

7.6.1.2 ArmA sub-scale unidimensionality and scaling properties

Unidimensionality was initially examined using Principal Components Analysis (PCA) and then further evaluated using Mokken analysis. Mokken analysis confirmed unidimensionality, with an H coefficient for the passive and active items and total sub-scales of greater than 0.3, as recommended by van Schuur (van Schuur 2003).

In the passive function sub-scale Principal Components Analysis (PCA) loading was consistent on this sub-scale (see Table 7.11; page 212). Following Mokken analysis a moderate unidimensional scale was identified conforming to monotone homogeneity. Further reduction of items to improve measurement properties was not considered at this stage, however possible violations of double monotonicity were identified in Item 6 (putting on a splint) and Item 7 (positioning the arm) of the passive function sub-scale. Removal of these items resulted in a strong unidimensional scale conforming to double monotonicity. However, given the preliminary nature of the evaluation and the limited sample size (n=92) for this analysis, items were not excluded at this stage.

During the selection of the ArmA items in Chapter 7, item 8 (active function) - 'Effect of arm on balance when walking' - was considered for inclusion in both sub-scales. Since walking is an active function it was included in the active function sub-scale but could have been included under passive function. This is because the influence of the arm on walking is usually passive in the form of increased spasticity or associated reaction due to the effort of walking. The upper limb can impact on walking and balance either because expression of spasticity increases due to effort and therefore alters the posture of the body, or because it acts as a weight with no active movement which can also alter posture.

Following rotation, PCA indicated this item, to be closest to the passive function sub-scale than the active function sub-scale (see Figure 7.7; page 221; item a81). This item was the third lowest performing item following Mokken analysis of the active function sub-scale (item 8; see Table 7.18; page 223). However, the active function sub-scale evaluation, was limited due to the ceiling effect in this sub-scale and will require further evaluation.

The ArmA is a measure of difficulty with passive and active function. This is more straightforward for the passive function sub-scale, where measurement is confined to the affected upper limb. However for the active function sub-scale some items are unimanual, other items are bimanual, and the impact of hemiplegia may change what is normal function for the patient. Although a reflective measure of difficulty with active function is still produced, the unaffected upper limb may carry out some tasks unilaterally, particularly if this limb is dominant. This will result in some items scored as having 'no difficulty' by being performed by the unaffected limb, while performance of the task by the affected limb is not possible. Measuring overall difficulty with active function provides a true indication of function, but individual items are not reflecting only the performance of the affected upper limb. An additional question has therefore been added to the active function questions to ask if the affected limb carries out each item or if it is done by both limbs (see Appendix 22). Understanding which limb is used when performing each item is useful from a clinical perspective for treatment planning rather than measuring overall difficulty with active function and may enable additional analysis for research. The instructions were not just changed to allow use of the affected limb only, because ArmA is a measure of function not just affected arm use.

Following this evaluation of the ArmA, one previously excluded passive function item has been identified which merits further consideration regarding its place in the measure. During item reduction, 'cleaning around the affected elbow' was removed during the first round of Delphi consultation. This item was removed on the recommendation of eight members of the consultation group because the item was possibly not fully understood. However, from a clinical perspective 'Ease of elbow crease hygiene' continued to be set as a goal for participants in the cohort study (n=6).

Based on this finding, consideration could be given to including this item in the passive function sub-scale and evaluating the scaling properties of the modified measure. This finding emphasises the need for a multi-modal approach to item reduction, which could have been strengthened by further referring to the goals of clinical intervention before psychometric testing.

Additional work might also be explored in the future to ensure that dimensionality and measurement structure of the ArmA remain consistent across time points rather than just the baseline time point. This has particular relevance if the measure is to be used for before and after intervention studies as is proposed with ArmA. Using a latent variable model (LVM) this could be addressed by examining longitudinal construct validity and one method to evaluate it would be reviewing the factorial structure of the ArmA at the different time points using confirmatory factor analysis. An alternative approach using IRT would be the application of Rasch analysis examining differential item functioning at each time point.

Another issue highlighted during the ArmA development and evaluation is the complexity of scoring the passive function sub-scale and the possible effects on dimensionality and validity. The passive function sub-scale may be scored by the patient alone (if they undertake all tasks), the patient and carer in combination (if tasks are undertaken in combination) or the carer alone (if only the carer is involved in carrying out the tasks). This presents difficulties because patient and carer information may be different and an alternative approach would be to collect two scores (patients and carers). However, in many instances both patient and carer are involved in carrying out the task and a single score to represent the tasks difficulty is desirable for analysis and comparison. A combined score may actually also be more representative of the manner in which the task was completed. The combining of the score could be considered to be creating a further dimension giving a possible three (patient completed, carer completed and combined completed). This issue deserves further exploration in future work.

In practice, in this thesis, a pragmatic approach has been taken with scoring of the ArmA by the person or persons carrying out the task. Use of a simple solution in this

manner is important for clinical utility, but future work is needed to consider the relationship between these different scores in greater depth. Possible methods could be confirmatory factor analysis and correlations between the two (or three) sets of scores (patient completed, carer completed and combined).

7.6.1.3 Internal consistency

Cronbach's alpha for the passive function and active function subscales showed high internal consistency. A theoretical risk of item redundancy was presented by the results for internal consistency in the active function sub-scale, with reference to the recommendations of Terwee and colleagues (Terwee et al. 2007). However, as referred to in the results this issue appeared to be related to the ceiling effect seen in the study group in the active function scale.

7.6.1.4 Test re-test reliability (Reproducibility)

Test–retest reliability after 24 hours used quadratic-weighted Kappa and produced a positive rating for reliability conforming to the recommendations of Terwee and colleagues (Terwee et al. 2007). Test re-test reliability was adequate with sufficient power to provide an indication of the reliability of the active and passive sub-scales.

An evaluation of the clinimetric properties of the MAL was undertaken by van der Lee and colleagues for the assessment of arm use in hemiparetic patients (van der Lee et al. 2004). The evaluation assessed both the 26-item and 14-item MAL but only the 14-item version results were reported. Internal consistency was evaluated using Cronbach's alpha and test re-test reliability was evaluated using a two week time interval during which no change was expected using the Bland and Altman 'limits of agreement method' (Bland and Altman 1986). The Bland and Altman method, unlike weighted Kappa used for ArmA, is designed as a measure of absolute agreement for interval data. A fundamental weakness in using this method is that it is a parametric approach, requiring interval level measurement, which has not been established for the MAL items.

Uswatte and colleagues examined MAL-14 reliability in two groups of patients in two separate studies evaluating CIMT (Uswatte et al. 2005). Cronbach's alpha was used as in ArmA evaluation, for internal consistency and indicated that MAL-14 was internally

consistent. Test re-test reliability was evaluated using Pearson correlation, which is considered by Terwee and colleagues to be insufficient because systematic differences are not taken into account. Pearson correlation will therefore usually produce ratings higher than actual reliability calculated using other methods. The evaluation of test re-test reliability in ArmA using the weighted Kappa coefficient is a more robust method of demonstrating test re-test reliability and is supported by both Terwee and colleagues and Streiner and Norman (Streiner and Norman 2003; Terwee et al. 2007).

7.6.1.5 Responsiveness to change

A significant difference was seen between responder and non-responder groups in ArmA passive function but not in ArmA active function nor in LASIS, Barthel Index or DASH. Preliminary effect sizes (ES) and standard response means (SRM) were calculated for the ArmA, to enable comparison with the other measures. Calculation of ES and SRM was undertaken with the acknowledged limitation of not having established interval scaling. The ESs and SRMs were greater for the ArmA sub-scales (passive 0.29 and 0.30 and active 0.21 and 0.16 respectively) than for the other measures. However, these results for the ArmA were still relatively low. The values for the ArmA active function sub-scale are explained by the lack of changes seen or expected in the study group as a whole. However the results for the passive function sub-scale are disappointing and require further exploration once interval scaling has been established using Rasch analysis for the ArmA in a larger study population.

The difference in responsiveness, (ES and SR) between the LASIS and the ArmA is surprising since they both contain a number of similar passive function items. However there are also different items in the ArmA and the LASIS. The LASIS total score also incorporated both the carer and the patient score in this analysis and may have led to confusion between patients and carers regarding who should be completing items, having already completed the ArmA in some cases. Patients and carers were not specifically asked about their understanding of LASIS completion so it is difficult to determine if this was the case. One issue with LASIS that may also have made a difference was asking the patient or carer if the particular item was difficult with a yes/no response before rating the item on the scale. This appeared in some instances to lead patients and carers to say that an item was difficult but then not completing the

scale resulting in missing data. Patients and carers may also have indicated that an item was not at all difficult, when in fact it was mildly difficult in some cases.

The approach taken of comparing change in the ArmA with an overall rating of response to intervention by clinicians is similar to that taken by other authors e.g. van der Lee et al. 2004. Responsiveness to change in MAL-14 was also evaluated by van der Lee and colleagues by comparing change with that in ARAT and a global rating of change (GRC) by patients after two weeks. No significant positive correlation was found between MAL-14 and ARAT or GRC. The approach taken in evaluation of responsiveness was similar in the ArmA with rating of response by clinicians compared with change demonstrated by the ArmA. The ArmA and LASIS (passive items) in BTX intervention, measure very similar functional issues, while MAL-14 and ARAT measure patient-reported every day function and clinic-performed functional tasks respectively. While a relationship should exist between every day function and functional tasks performed in a clinic, differences will also be present (as discussed in Chapter 1), which may account for the lack of correlation observed in the van der Lee study (van der Lee et al. 2004).

Responsiveness of the ArmA measure has been evaluated for the passive function sub-scale but evaluation of the active function sub-scale is needed in a patient group showing change in this domain. Qualified support is provided for passive function responsiveness, but this requires further evaluation in future work.

7.6.1.6 Interpretability

A significant difference was identified between responders and non-responders for the passive function sub-scale and therefore MIC was calculated. However, interval scaling has not been demonstrated for the ArmA and will need further evaluation in future stages of development and evaluation. A change of two or more points on ArmA, was deemed a preliminary indication of clinical significance taking into account calculations of MIC and evaluation of sensitivity and specificity. Absence of change in the active function sub-scale was expected once initial data collection had begun, due to the lack of active function goals for intervention in this sample.

The establishment of MIC for the passive function sub-scale is important in using the ArmA in clinical practice to indicate meaningful change following intervention and enabling decisions about clinical effectiveness. The MIC will need to be further explored in the ArmA once interval scaling properties have been established using Rasch analysis before a figure can be effectively applied.

7.6.1.7 Feasibility

If measures are to be applied in clinical practice they must be practical to apply in the setting in which they will be used. Feasibility is concerned with ensuring that outcome measures can be both practical to use in routine practice and retain their psychometric properties (Slade 2002b).

Some limitations arise from assessing feasibility as a component of a psychometric study. A psychometric study uses a design to allow evaluation of the measure or includes this evaluation within another research study, while feasibility by definition should be evaluated in normal clinical practice. The ArmA evaluation required a research design but this was applied in normal clinical practice situations for spasticity intervention. Patients were receiving standard treatment and the evaluation of feasibility in this context was adequate. However, limitations do arise from this compromise between normal clinical practice and the need to present the ArmA alongside a number of other measures in the questionnaire pack. Feasibility evaluation in a non-experimental patient group would confirm the current feasibility findings. More detailed evaluation of feasibility could have been useful, but given the burden of the questionnaire pack (including the ArmA, LASIS, DASH, Barthel Index and the feasibility questionnaire), it was decided not to increase the size and complexity of the feasibility questionnaire further.

The psychometric evaluation of the ArmA has been discussed and compared with findings for MAL, ABILHAND, LASIS and DASH. Table 7.35 provides a summary of the quality of the psychometric properties of these four (including four versions of MAL) measures.

Table 7.35 Comparison of psychometric evaluation for the ArmA with the MAL, ABILHAND, LASIS and the DASH.

Measure	Time	Admin. Burden	Content Validity	Internal Consistency	Construct Validity	Floor / Ceiling Effect	Reliability	Agreement	Responsiveness	Interpretability
ArmA	+	+	+	+	+	+ (passive) - (active)	+	+	±	±
MAL-14	-	+	?	+	±	±	-	-	-	±
MAL-26	-	+	?	+	±	±	-	?	-	?
MAL-28	-	+	?	+	±	±	±	±	?	?
MAL-12	±	+	?	?	?	?	?	?	?	?
ABILHAND	-	+	+	+	+	-	+	+	+	+
LASIS	+	±	?	?	?	?	?	?	?	?
DASH	-	+	+	+	+	-	+	+	+	+

Method or result was rated as: + Adequate; ± Unclear; - poor;? no data available.

Following development and psychometric testing, the ArmA preliminary findings have been inserted into the theoretical hierarchy of measurement items produced following the systematic review. The insertion of the ArmA into this hierarchy enables comparison of items with those in other measures. The ArmA measure items overlap with the LASIS and include many of the same passive function items. It also has items in common with all four versions of the MAL and has some items in common with the ABILHAND, but does not cover the most complex upper limb tasks (see Table 7.36).

Table 7.36 Comparison of items in the ArmA with those from the systematic review

Functional Items	LASIS	ArmA	MAL -14	MAL -26	MAL -28	MAL -12	ABIL-HAND
Passive Function Items							
Cleaning the palm affected hand	1	1					
Cutting fingernails affected hand	2	2		25*			4*
Cleaning the affected elbow	3						
Cleaning the affected armpit	4	4					
Cleaning the unaffected elbow	5						
Putting arm through coat sleeve	6	5	1*	1*			
Difficulty putting on a glove	7	3					
Doing physiotherapy exercises to arm	9						
Put on a splint		6					
Position affected arm comfortably		7					
Active Function Items							
Difficulty rolling over in bed	8						
Difficulty balancing standing	10						
Difficulty balancing walking	11	15					
Hold object steady, use other hand (jar[a])	12	12[a] 14					10[a]
Steady myself while standing			2	2			
Carry an object from place to place			3	3	23	12	
Pick up fork or spoon, use for eating		13	4	4	24	10	
Comb hair		18	5	5	25		
Pick up cup by handle			6	6	26	11	
Hand craft/card playing			7	7			
Hold a book for reading			8	8			
Use towel to dry face or other body part			9	9			
Pick up a glass		9	10	10	20	5	
Pick up toothbrush and brush teeth		19	11	11	21	6	
Shaving / make-up			12	12			
Use a key to open a door		10	13	13	22	7	
Letter writing/typing		11	14	14		8	
Poor coffee / tea				15			
Peel fruit or potatoes				16			3
Dial number on the phone		16		17			
Open / close a window				18			
Open an envelope				19			
Take money out of a wallet or purse				20			
Undo buttons on clothing				21			
Buttons on clothing (shirt[a], trousers[b])		8		22	27[a]		13[a] 17[b]
Undo a zip				23			
Do up a zip (jacket [a], trousers [b])				24			11[a] 21[b]
Other optional activity				26			

Functional Items	LASIS	ArmA	MAL -14	MAL -26	MAL -28	MAL -12	ABIL-HAND
Tuck in a shirt/blouse		17					
Drink from a cup or mug		20					
Turn on a light with a light switch					1		
Open a drawer					2		
Remove item of clothing from drawer					3		
Pick up phone					4	1	
Wipe kitchen counter					5		
Get out of car					6		
Open refrigerator					7		
Open a door by turning a door knob					8	2	
Use a TV remote control					9		
Wash your hands					10		
Turn water on/off with faucet					11	4	
Dry your hands					12		
Put on your socks					13		
Take off your socks					14		
Put on your shoes					15		
Take off your shoes					16		
Get up from chair with arm rests					17		
Pull chair away from table before sitting					18		
Pull chair toward table after sitting					19		
Eat half a sandwich or finger food					28	3	
Use removable computer storage						9	
Hammer a nail							1
Thread a needle							2
Wrap gifts							5
File nails							6
Cut meat							7
Peel onions							8
Shell hazel nuts							9
Open pack of chips (crisps)							12
Sharpen pencil							14
Spread butter							15
Fasten ‘snap’ (press-stud)							16
Take the cap off a bottle							18
Open mail (post)							19
Squeeze toothpaste							20
Unwrap chocolate							22
Wash hands							23

Key:

Items in the table are given the number at which they appear in order in the measure.

Items in the LASIS and the ArmA included under passive function all asked respondents ‘how difficult’ a task was to undertake related to care of the limb by the patient him or herself or a career.

* Items in the passive function section included in MAL-14, MAL-26 or ABILHAND could be done either passively or with more active involvement by the individual, with the focus being on active involvement in these measures.

[a] and [b] refer to specific objects used for the functional items in a measure.

7.6.2 Evaluation of functional change: A cohort study

The application of the ArmA during the cohort study identified a small but significant improvement in passive function at 8 weeks, which was maintained at 16 weeks (see Table 7.33; page 237). Change in passive function was also confirmed by LASIS. Passive function change corresponded with an initial decrease from baseline in spasticity recorded by MAS by 8 weeks followed by an increase in spasticity by 16 weeks. The reason for the implied maintenance of functional benefit in the presence of an increase in spasticity is not entirely clear, although the most plausible explanation is that PT intervention used in conjunction with BTX may be responsible. Two other possible reasons for the results include; a) natural improvement over the period of BTX action in spasticity or b) inaccuracies in measurement of either function or spasticity. It is unlikely that natural improvement in spasticity would occur over the 16-week period of the study, because this is a relatively short period in which to expect spontaneous improvement in spasticity presentation, which is not acute. Natural improvement also does not explain the re-increase in spasticity (measured by MAS) seen at 16 weeks. Measurement inaccuracy is a possibility, but is also unlikely given the consistent pattern produced in a majority of participants and the concurrence between the ArmA and LASIS in function.

The study has also demonstrated that the ArmA is able to identify change following intervention, which was not detected by the other measures used with the exception of LASIS for the overall analysis. The LASIS passive function items demonstrated change when the Friedman test was applied overall, but not for the individual analysis at 8 and 16 week outcome points. The ArmA did however show significant differences between baseline and 8 and 16 weeks as well as differentiating responders and non-responders in the responsiveness evaluation unlike the LASIS. These findings provide modest support for the use of the ArmA as a more responsive measure following focal spasticity intervention, than the other measures applied. Utility of the ArmA seemed acceptable, in this preliminary evaluation, and in the context of being able to detect change, its further evaluation is supported.

The results of the cohort study also seem to indicate a trend to greater improvement in passive function towards 16 weeks from that seen at 8 weeks. This pattern of improvement has been observed by other authors, such as Francis and colleagues, who identified maximal improvements in passive function occurring at some time after maximal improvements in spasticity in some patients (Francis et al. 2004). One possible reason for the delayed improvement in passive function is the need for time for the PT interventions to take effect.

7.6.2.1 Significance of passive function improvement

Other authors have also identified sustained improvement in passive function following treatment with BTX and PT. In a recent study by Shaw and colleagues a combination of BTX and PT interventions were applied (Shaw et al. 2010). The experimental group received BTX and PT. The control group received PT alone. PT consisted of a stretching programme for all participants and a task-training programme for participants with an Action Research Arm Test (ARAT) score above four. Spasticity was reduced in the intervention group at one month, but not 3 or 12 months compared to the control group. Differences in active function improvement between the groups were not identified. However, improvements were identified in individual passive function tasks at one, three and 12 months. These results support the findings of the current cohort study in identifying a pattern of maintained improvement following BTX with the application of a PT maintenance programme. Passive function improvement was not the primary outcome in this study and these improvements were therefore not emphasised in the findings. Nevertheless improvements in passive function were significant and the focus on active function in this study illustrates the over emphasis active function improvement often receives in the literature and practice.

7.6.2.2 Physical therapy interventions

Specific evaluation of PT intervention in combination with BTX has been limited (see Chapter 1; section 1.3), and therefore comparisons of the cohort study with other work are restricted. The following two studies are however briefly discussed in relation to the findings.

Giovannelli and colleagues, in their randomised controlled trial with multiple sclerosis patients, identified that improvements in Modified Ashworth scores were significantly

greater in patients receiving physiotherapy than those who were not (Giovannelli et al. 2007). Unfortunately, the actual PT interventions are not specified in the reporting of this study. The results from Giovannelli and colleagues have similarities to the findings in the cohort study, with spasticity maintained in their work until 12 weeks in the experimental group and in this cohort study until 16 weeks.

Carda and colleagues undertook a case control study in patients who all had BTX intervention, investigating application of adhesive strapping of the wrist and fingers compared with electrical stimulation combined with a resting splint (Carda and Molteni 2005). This study again demonstrated a similar pattern of change in spasticity to that identified in this thesis. Since Giovannelli et al (2007) and Carda and Molteni (2005) do not record changes in any measure of function, the timing of change in spasticity and in function cannot be compared.

7.6.2.3 Recording of physical therapy intervention

The study has also enabled the initial development of a system for recording PT interventions in a clinical practice setting for focal spasticity management in the upper limb. This preliminary system of recording could now be formally refined, evaluated and compared to other similar classification systems developed, for example that by Donaldson and colleagues (Hunter et al. 2006; Donaldson 2007; Donaldson et al. 2009) and De Wit (De Wit et al. 2007). However the complexity of constructing and developing treatment schedules should not be underestimated as emphasised in the work of DeJong and colleagues (DeJong et al. 2004; DeJong et al. 2005; Gassaway et al. 2005) and emphasised by Jette in a commentary on the conclusions drawn from the work (Jette 2005). The recommendation of developing a treatment schedule echoes that by the recently published botulinum toxin: international consensus statement (Sheean et al. 2010). Further strengths, limitations and implications of the cohort study are discussed in Chapter 8.

7.6.3 ArmA evaluation strengths and limitations

Limitations in the study were;

a) ceiling effects in the study population for active function;
b) sample for psychometric evaluation
c) feasibility evaluation within a psychometric study.

However important strengths have been identified in;

a) the resulting measures evaluation of passive function;
b) the initial evaluation of feasibility and
c) the potentially low burden for patients of completing the measure.

7.6.3.1 Ceiling effects for active function

The active function sub-scale could not be fully evaluated because of the limited changes in active function occurring in the main study group (Group 1). This necessitated the addition of Group 2 to evaluate test-retest reliability and internal consistency of the active function sub-scale. Ceiling effects may be a characteristic of the scale or may be a result of targeting the scale at a population, which is unlikely to change in the domain measured. The ArmA has two sub-scales designed to address the second of these potential problems, with the knowledge that the majority of patients make passive function improvements, but a small minority also improve in active function (McCrory et al. 2009).

The study population as a whole had very limited arm function, therefore ceiling effects occurred in the active function sub-scale. Possible reasons for this ceiling effect might be that items in the active function sub-scale were too difficult and that all the items have a similar level of difficulty. However, these explanations do not fit with the other information collected. Goal setting for the majority of patients did not include active function goals indicating that improvement in active function for these patients was unlikely. In addition, the clinical examination recorded for these patients did not record improvements in active function or selective movement following intervention. As

these patients have very limited active function, the ceiling effect does not therefore necessarily indicate a deficit in the measure.

Conceivably, if patients were very able a floor effect might be seen for difficulty in passive function with none of the passive function tasks being perceived as at all difficult (i.e. ceiling effect in passive function). These actual ceiling and possible floor effects are the reason for the construction of the ArmA in two distinct sub-scales to capture this information.

It might be argued, that given the very small number of participants improving in active function, measurement of active function following spasticity management intervention is not appropriate. However, this would lead to an inability to detect improvement in active function in the small number of patients who do improve in this dimension. Improvements in active function are multi-factorial, involving not only management of focal spasticity, but also for example task specific training. Nevertheless focal spasticity intervention may be a component in the overall improvement seen in a small minority of patients (rather than passive function improvement in the majority undergoing spasticity intervention) and recording this is therefore of clinical importance.

7.6.3.2 Sample

An overall limitation of the psychometric evaluation, was the sample size. The sample size achieved (n=92), was smaller than the target of 100 participants identified at the outset of the study. As discussed in section 7.3.1 (page 191), sample size for PCA is not clear. However, the target for sample size based on currently available evidence was not met. Future work should focus on increasing the participant numbers used for these evaluations to confirm the preliminary findings. Despite this limitation, the passive function sub-scale psychometric properties have been evaluated as planned at the start of the study.

The psychometric evaluation group (as with the goal setting analysis), were younger than the common age range associated with a general stroke population (mean age 44.5). However, these mean ages did reflect the age range of patients seen in specialist

rehabilitation services and the mean age of patients from the two services used for recruitment. The age range in the development and evaluation groups used may reflect a theoretical limitation. The practical implications however, are likely to be limited due to the significant disability in the psychometric study group (Barthel Index 7) and inclusion of patients with both significant communication issues and cognitive impairment, which are likely to be more important factors than chronological age.

7.6.3.3 Feasibility

Feasibility in this evaluation was focused on patients and carers who are the users of the measure. Patients and carers were prioritised as the most important and appropriate sources of data. Although the views of clinicians were obtained during the ArmA development and were used extensively in item reduction and confirmation. They were not sought again formally during the testing of the ArmA. Views from professionals may have been valuable in considering the impact of the ArmA in practice from the clinician's perspective. These data are currently being collected in the ongoing evaluation of the ArmA in clinical practice.

7.6.3.4 Low burden of ArmA for patients, carers and clinicians

In busy clinical environments, clinicians may be concerned that an additional outcome measure will add to their workload. Previous work in patient reported outcome measures (Greenhalgh et al. 2005) and standardised measures in general practice (Meadows et al. 1998) have emphasised that introduction of outcome measures into practice require clinicians to have ownership of the use of such measures. These studies also emphasise the need for measures to be feasible for use in clinical practice settings to ensure that they are acceptable to clinicians. This is where patient reported outcome measures, such as the ArmA, have a potential strength by having a low impact on clinician time. In the ArmA, the burden on patient and carers is also low. This should ensure that it can be applied in routine practice for spasticity management as well as research. In a previous survey of practice, routine use of outcome measures by clinical teams for spasticity management was shown to be variable (Turner-Stokes 2009a). Methods for improving use (such as the use of integrated care pathways) and self-report measures such as the ArmA, need to be considered.

7.7 Conclusions

Based on this evaluation of the ArmA, psychometric properties of the passive function sub-scale were generally supported. Further evaluation is required to confirm double monotonicity for ordinal scaling and to undertake Rasch analysis for interval scaling properties. In light of the scaling limitations of this evaluation of the ArmA, further work to examine responsiveness and MIC will be needed for the passive function sub-scale once fundamental measurement properties have been demonstrated.

The active function sub-scale suffered from a ceiling effect in this group of patients with generally severe spasticity and disability. Further evaluation of this sub-scale in a different group is indicated to evaluate all measurement properties, but in particular the scaling ability (ordinal and then interval) of this measure. Evaluation may be needed, in particular for responsiveness, in a non-spasticity intervention group to gain sufficient data to fully evaluate the active function sub-scale.

The ArmA can be completed by self-report as well as structured interview, making it useful for completion and return by post following clinic visits with a potentially low burden on clinician's time. The passive function sub-scale is now appropriate for preliminary application in clinical practice, however the active function sub-scale needs further evaluation before it is applied more widely.

Chapter 8 Thesis discussion, future work, and conclusions

8.1 Summary of findings

The major components of discussion for the different strands of this thesis (systematic review and patient identified items, ArmA development, psychometric testing and cohort study) have been located in the relevant chapters. This final chapter offers an overview of the key findings, general strengths and limitations of the thesis as a whole as well as methods and analytical techniques. Plans for future research and possible methods for addressing the limitations in this thesis are also proposed.

The clinical challenge leading to the development of this thesis involved re-conceptualisation of measurement of upper limb function following spasticity management from the perspective of patients and carers. A challenge arises from the need to measure passive and active function particularly in the context of everyday activities of relevance to focal spasticity intervention in the upper limb. Self-report measures are a means of capturing the impact of intervention from the perspective of the individual in the context of their normal lives and also offer the possibility to obtain follow-up information at a distance.

This thesis has made the following three key contributions to knowledge:

Measurement

The complexities of measurement have been explored in the context of variables that are not directly observable. The need to establish clear dimensions as a basis for measurement has been addressed in this work and the clinically conceived dimensions of active and passive function have been supported. Ordinal scaling of the ArmA passive function sub-scale has been partly evaluated, with further work identified in terms of ordinal scaling and in due course interval scaling using the Rasch analysis method.

Additive conjoint measurement was discussed in this thesis and the need for measurement tools of latent variables to meet these requirements was established. However, the application of classical test theory approaches, latent variable methods

and an item response method to establish ordinal scaling, have been identified as useful preliminary steps in measure development and evaluation. These methods result in a measure with partially established psychometric properties and a clear indication of current limitations and need for future work.

Active and passive function

The ArmA is a two-dimensional tool with two sub-scales evaluating active and passive function in the context of focal spasticity intervention. The concept of difficulty with upper limb functional tasks has been applied in a self-report tool, capturing the patient and carer perception of upper limb passive and active function limitation due to spasticity. While the concept of difficulty has introduced challenges, it has been successful in the central aim of recording a patient and carer perspective. The combination of patient and carer scoring also has challenges, but for passive function has been effective in reflecting the impact for the combined patient/carer unit.

The primary focus of measure development was the passive function sub-scale, which received the major theoretical focus. Passive function has received limited attention in the literature and yet is important to evaluate in patients with hemiparesis undergoing focal spasticity intervention in the upper limb. Passive function was expected to be the main area of improvement in the study group undergoing spasticity management interventions.

Maintenance of passive function improvement

The application of the ArmA during the cohort study identified that passive function was maintained from 8 weeks to 16 weeks post injection with BTX. The reason for maintenance of functional benefit in the presence of increasing spasticity is not entirely clear. One possibility is the positive effect of the physical therapy (PT) interventions provided in conjunction with BTX. Other authors have identified changes in passive function maintained to 16 weeks and in one study up to one year following BTX intervention (Shaw et al. 2010). It has also been suggested that, in some patients, maximal improvement in passive function may develop some time after maximal improvement in spasticity specifically following BTX intervention (Francis et al. 2004). There has been a focus on active function improvement in both clinical practice and

research (see Shaw et al 2010). The cohort study has provided additional evidence that it is largely passive function improvements that are seen following focal spasticity intervention including BTX and PT.

8.2 Strengths and challenges

Overall, the work in this thesis has major strengths in addressing the challenges of a different approach to measurement of functional outcome in the hemiparetic arm following focal spasticity intervention. The approach has been to re-conceptualise activity into passive and active function and evaluate against overall difficulty of everyday performance. The following section (8.2.1) will initially discuss the strengths and challenges related to the methods and analytical techniques used in this thesis. Then in Section 8.2.2 the strengths and challenges that relate to the findings will be discussed.

8.2.1 Strengths and challenges of the methods and analysis

A systematic approach was used in developing the ArmA by applying a structured method of combining findings from the literature with findings from clinical practice (Chapters 4 and 5). Consensus techniques were used to combine and confirm literature and practice findings in the selection of items. A psychometric evaluation was subsequently applied to the resulting measure; evaluating distinct properties and using classical test theory as well as latent variable and item response techniques.

8.2.1.1 General strengths

The multi-methods approach, evidenced above, is a strength of this thesis and has allowed cross-confirmation of measurement properties giving increased rigour to the process. A further strength comes from the use of reported 'difficulty' in passive function as a model for outcome evaluation. The theoretical approach of recording task-difficulty in function has been used in contrast to some existing measures, but in common with others such as DASH or LASIS. Many measures administered at clinic appointments assess standardised versions of functional tasks (e.g. the Action Research Arm Test (Koh et al. 2006)) and, in some measures, items assessing function are mixed with those assessing impairments (e.g. the Motor Club Assessment (Wade 1992b) page 147).

In the ArmA rather than an observer assessing individual task performance, patients (and/or carers) are asked to rate difficulty. This approach has been used by other authors in the field e.g. (Bhakta et al. 2000a) and has particular relevance to passive function. Passive function tasks are not necessarily performed by the patient, but can be performed by patient or carer. As tasks are not active, difficulty in task performance is a particularly relevant concept.

8.2.1.2 General challenges

However, a challenge in this thesis has been evaluation of difficulty in active function. While the concept of difficulty works well in passive function, it does not apply as easily to active function, although it has been used in active function measures such as ABILHAND (Penta et al. 2001) and MAL (Uswatte et al. 2005). When specifically related to the affected upper limb, measurement of active function may be complicated by both performance of bimanual and unimanual tasks and hand dominance (Jones 1990; Alon et al. 2003). A development following the psychometric evaluation is to record difficulty in active function for all items regardless of which limb is used (affected or unaffected). Patients and carers are then asked to indicate for each task the limb used or that it was undertaken using both limbs. The revised version of the ArmA is presented in Appendix 22.

The modification to scoring the active function sub-scale should make the ArmA conceptually easier for patients and carers to complete. It is still possible to review how individual items change even if this involves a change of limb completing the task, for example as an affected dominant limb improves in active function. Further evaluation will be required of the revised measure.

Another challenge identified in Chapter 7 (Section 7.6.1.2; page 243-246), was the complexity of scoring the passive function sub-scale and possible effects on dimensionality and validity. The passive function sub-scale may be scored by the patient alone, the patient and carer in combination or the carer alone (if only the carer is involved in carrying out the tasks). This presents difficulties because patient and carer information may be different. However, in many instances both patient and carer are involved in carrying out the task and a single score to represent the tasks difficulty was

decided upon because it more fully represented reality for patients and carers. This issue deserves further exploration in future work.

8.2.1.3 Strengths and challenges of development and psychometric evaluation

The following section will discuss the strengths and challenges encountered during the development of the ArmA and its psychometric evaluation.

Item Generation

The use of the literature to generate items for possible inclusion in the ArmA, has strengths in utilizing the work of others evaluating upper limb function. Items identified in this way are more likely to have face and construct validity. However, as identified in the systematic review, all of the existing measures, excluding LASIS, address only active function. This process is therefore unlikely to identify many passive function items. The review of goals for focal upper limb spasticity intervention aimed to redress this imbalance because mainly passive function goals are set in this area.

The use of the systematic review produced a large pool of active function items. However this approach risked reflecting the accepted wisdom and in particular may be prone to reflect the items identified by clinicians rather than patients and carers (Lomas et al. 1987). Much discussion has been undertaken related to active function improvement specifically following BTX administration for focal spasticity (Bakheit 2004a; Bergfeldt et al. 2006; Shaw et al. 2010). However to date, improvements have generally been in passive function, with only a small minority of patients showing active function improvement with focal spasticity intervention including BTX (Shaw et al. 2010). While measuring active function still seems important for detecting change in those patients who do make improvements in this area, passive function improvement is most relevant to measure for the majority of participants.

The goal-setting review identified only two additional items, which were both passive function. However, it enabled the confirmation of other items identified from the literature; all but one of these were passive function. This analysis included goals set with a relatively small group (n=16) and could have been expanded further to explore

other items set as goals in this area. To address the lack of active function goals, purposive sampling could be employed to identify more active function items, possibly in participants undergoing different focal interventions. However, if applied in spasticity intervention, given that active function goals (as demonstrated by both the goal analysis and cohort study in this thesis) are set so infrequently, even this approach is unlikely to identify a large number of relevant new items.

Other methods of item generation could have been applied, such as focus groups or interviews and have been discussed in Chapter 3. However, the advantage of the goal setting analysis, came from focusing on the particular area of investigation, avoiding the problems associated with group interactions in eliciting patient and carer perspectives. Disadvantages include the clinician potentially influencing the patient in the goal setting process (Reed and Roskell-Payton 1997), and lack of reflection of a larger patient group in considering items (goals may be specific to that individual). The sample reflected the patient population in specialist neurological rehabilitation services. However, it may not be as relevant to other groups (e.g. an older post stroke population), and further evaluation will be required to determine applicability of the ArmA in other populations in which it may be used. The restriction of goals to a very specific area for upper limb spasticity management, may have limited identification of active function items in particular. However, this is balanced against the limited active function goals set in this area and the large number of active function items identified from the literature.

Passive function items were identified in the literature, but only from one measure, the LASIS. The goal setting analysis identified additional items, but only two. This may result in very similar items for the passive function sub-scale, which may be more appropriate to the formation of a clinical 'check-list' rather than a measure due to the possibility of similar difficulty of the items. A related issue is that clinical presentation of spasticity may vary between individuals and therefore the items that change following intervention may also vary. This raises the issue of whether the ArmA is a clinical 'check-list' rather than a measurement scale. While an individual patient's clinical presentation can certainly vary, consistent problems do arise as evidenced in the goal-setting analysis and the subsequent goals set in the cohort study. The item generation methods used have attempted to obtain the important items relevant to

passive function. Other methods of item generation, as already discussed, could have added to this process for passive function. Nevertheless, passive function items have been explored in some detail for dimensionality using two methods and the initial indication from preliminary Mokken analysis for scaling is that an ordinal scale is apparent.

Item reduction – Delphi consultation

Item reduction may take different forms, but in this thesis Delphi Consultation was used to obtain opinions of experts to identify the most important items reflecting the construct. The Delphi consultation (three rounds of consultation) group consisted of clinicians and was then followed by a wider consultation to confirm the findings with a different and larger group of clinicians as well as patients and carers. As discussed in Chapter 6 a possible limitation of prioritising the items using the Delphi process and wider consultation is that a set of items with a similar degree of difficulty may be the result, limiting the range of the scale. While items measuring the same construct or dimension are desirable, if the items also have similar difficulty, a potential limitation to the breadth of the hierarchical scale may occur. In this situation, items instead of forming a wide scale may cluster around one particular area of the dimension. However, preliminary evaluation of the ArmA indicates that it forms an ordinal scale. Passive function items for the upper limb also evaluate a relatively focused issue and therefore concerns over the breadth of the scale are likely to be more theoretical than an actual limitation.

Ceiling effect

In this thesis, for active function, the focus was on identification of items, which had low to moderate difficulty with fewer higher function items. A ceiling effect was observed in the active function sub-scale during the evaluation. One interpretation of this finding is that the active function items are still too difficult for the population tested (i.e. easier items may have detected change). However, given that active function goals were set in a very limited number of participants, a more plausible explanation is that active function was not possible for the majority of the study group (i.e. the majority of participants were unable to perform active function).

The probability that the ceiling effect is a feature of the study population raises the question of whether the ArmA active function sub-scale is 'well-targeted'. Linacre has argued that if many patients are at the margins of a scale, then the measure should be considered not well targeted (Linacre 1994). The implication is that either a larger study sample will be required to confirm measurement properties or better targeting of the measure to the sample will be needed to evaluate the measure. Pallant and Tennant (2006) have emphasised the need, particularly in clinical practice, to ensure that measures are appropriately targeted at the population being assessed. Good targeting is therefore seen as important for good measurement, as well as evaluation of such measures (Hagquist et al. 2009). However, the very reason for construction of the ArmA with two sub-scales is to ensure that items are appropriately targeted at passive function, but that active function changes are recorded if they occur. This may in part explain why, despite the lack of active function change in the study population, 74% of patients and carers valued the inclusion of the active function sub-scale.

As already discussed in Chapters 1 and 7, while for the large majority of patients only passive function change will occur; for a small minority active function changes also take place. A small, but not insignificant, proportion of patients (n=4) did in fact change in active function, which was detected by the ArmA. This finding both supports the need to measure active function and the utility of the ArmA in detecting these changes when they occur. In effect, assessment of active function in a spasticity intervention group is acting as a screening tool to identify those patients who are able to perform active function from the large majority who cannot.

Construct Validity

While correlations between the ArmA, LASIS and DASH were as expected, the positive correlations for passive function were moderate and higher correlations might have been expected between LASIS and ArmA passive function items. However, some difference in items occurs between the ArmA and LASIS and the scoring methods differ, both of which may in part account for differences between the scores. Despite correlations not being as strong as expected, the construct validity of the ArmA is supported, but would benefit from further evaluation in future work. Further evaluation might involve

longitudinal construct validity using principal components analysis (PCA) as well as correlations with LASIS and DASH at the other time points in addition to baseline.

The cohort study demonstrated that the passive function sub-scale of the ArmA showed significant change between baseline and outcome at 8 and 16 weeks, which was also shown to a lesser extent by the LASIS passive function items using the Friedman test. This gives preliminary support to the longitudinal construct validity of the ArmA passive function sub-scale.

Dimensionality

Principal components analysis was used initially to give an indication of the key constructs underlying the ArmA sub-scales. The aim of PCA is to reduce the dimensionality of a data set consisting of a large number of interrelated items (Jolliffe 2002). A possible criticism could arise with the use of PCA as opposed to exploratory factor analysis (EFA) in terms of the appropriateness of this method. While the purpose of EFA is to explore relationships between items, the purpose of PCA is to reduce dimensionality in the data. The purpose of PCA and EFA may seem contrary to each other, but in effect the two methods are performing a similar function. In this evaluation of the ArmA, where measure development has been based around two sub-scales, the use of PCA is an appropriate method for exploring the dimensions within the measure.

The method of PCA has also been applied in other studies examining the dimensions contained in rehabilitation measures (Bedard et al. 2001; O'Rourke and Tuokko 2003; Siegert et al. 2010). As in this thesis, some studies have also combined this approach with other methods of confirming findings, such as Mokken analysis or confirmatory factor analysis (Siegert et al. 2010). The combination of methods, using both a LVM and IRT approach, lends strength to the findings.

Scaling

The aim when developing a scale is to enable the differentiation between people using the sum of the items as a summary statistic. Differentiation and consistent ordering of items is also desirable to infer directly from a participant's score, their level of function

in practice. In this thesis, Mokken analysis using monotone homogeneity was explored, however double monotonicity only received exploratory investigation. Double monotonicity was not the main focus of evaluation because of the limited number of participants included in the psychometric evaluation; therefore expansion of the numbers included in this evaluation would have been valuable. However, the current analysis has provided an indication of ordinal scaling properties for the passive function sub-scale, which will be explored further in subsequent evaluation of the measure.

The preliminary evaluation of double monotonicity undertaken, indicated that two items could be removed to improve upon both the dimensionality and scaling properties of the passive function sub-scale. There is a possible dilemma from a clinical perspective, where specific items give useful clinical information, but do not 'fit' the construct and scaling demonstrated by IRT approaches. However, if they do not fit the construct or fail to demonstrate scaling properties, in the context of the whole scale or sub-scale, they cannot form part of the measure (also see Section 3.3.3.5; page 107). An alternative, suggested in chapter 3, was the retention of such items to inform the clinical picture if they provide particularly important clinical information, but not to form part of the measurement scale.

Another related issue, is the possibility that the ArmA could form a checklist rather than a measure. This would mean that important clinical items (e.g. axillary hygiene) are recorded as single items solely for the specific clinical information they provide. In the case of the ArmA passive function, preliminary scaling properties have been demonstrated and although following further evaluation, some items might be excluded; measurement with the remaining items seems possible. Both passive and active function sub-scales require further evaluation and in the passive function sub-scale it is still possible, given an increase in sample size, that removal of items will not be required. Nevertheless, the ArmA appears to form an ordinal scale for passive function based on current evaluation.

Reliability (Reproducibility)

In this evaluation of test-retest reliability, the time interval between self-completion of the ArmA at baseline and completion 24 hours later with postal return was relatively

short. However, it is possible that patients and carers had some memory of completion of the ArmA the day before. It may have been appropriate to increase the time between the first and second completion of the ArmA to reduce recall from previous completion. However, increasing the time interval may also have resulted in a poorer questionnaire return.

The ArmA was completed at the 'clinic' in the first instance and then 'at home' and returned by post. This could constitute completion under different conditions, introducing a potential confound for test-retest reliability. However, as completion was by self-report on both occasions this was minimised.

Responsiveness

In the psychometric evaluation, a significant difference was seen between responder and non-responder groups in ArmA passive function, but not in ArmA active function nor in the LASIS, Barthel Index or DASH. Therefore, a significant limitation in the evaluation of responsiveness for the active function sub-scale was the minimal change in active function seen in the study group, which prevented full evaluation of this sub-scale and active function responsiveness in particular. However, strengths are seen in the ArmA passive function sub-scale detecting a response when the other measures were unable to show this.

The ES and SRM, as already discussed (Chapter 7; Section 7.6.1.5; page 247) are based on parametric assumptions and are therefore preliminary as evaluated for an ordinal scale in this thesis. However, these methods were used in common with the evaluation of many other scales to allow comparison. The ES and SRM were larger for the ArmA passive function sub-scale than other measures used, but were still moderate, which has also been discussed in Section 7.6.1.5.

Interpretability

Calculation of MIC using both the criterion-based method and distribution-based method, use mean and standard deviation and are therefore parametric methods. As already discussed in Chapter 7; Section 7.6.1.6 (page 248), this is a significant limitation that needs to be addressed in future work. These methods arrive at a single

figure to indicate clinically meaningful change across the whole of the scale. Interval scaling is therefore required to ensure that the amount of change that is clinically meaningful in the middle of the scale is also meaningful at the extreme ends of the scale. Further work to evaluate the MIC will be needed on an interval version of the ArmA following Rasch analysis.

Feasibility

Limitations are apparent in the evaluation of feasibility because of the compromise between normal clinical practice and the need to present the ArmA alongside a number of other measures in the questionnaire pack. As already discussed in Chapter 7; section 7.6.1.7 (page 249), feasibility evaluation in a non-experimental patient group would confirm the current feasibility findings. However, the psychometric evaluation was undertaken, as far as was possible, in a normal clinical practice setting and provides a strong indication of feasibility of the ArmA scale.

8.2.1.4 Cohort study

The cohort study using the ArmA revealed that passive function change occurs as expected following spasticity intervention and, importantly, that this is maintained as the direct effects of BTX subside. This finding corresponds with findings from other authors such as Bhakta and colleagues (Bhakta et al. 2000a) and Francis and colleagues (Francis et al. 2004). However, the small number of data collection points limits the resulting model of the relationship between change in spasticity and change in passive function. This finding is therefore hypothesis generating rather than confirmatory. Section 8.3 includes proposed further research to strengthen this finding.

The use of the International Classification of Functioning, disability and health (ICF) as a framework in the ArmA development ensures it maps onto the ICF. The goals of intervention set during the cohort study in this thesis, are presented in summary, classified according to the ICF in Table 8.1.

Table 8.1 Classification of cohort study GAS goals to World Health Organisation ICF codes.

Domain	Goal Area	Chapter	Sub-category 1	Sub-category 2
Body Functions				
	Pain reduction	2 – Sensory & Pain	b280 – Pain	
	Enable passive range of movement	7 – Neuro-musculoskeletal	b735 - Muscle Tone, b710 - Mobility of joints	
Activities				
	Balance improvement during walking	4 - Mobility	d450 - Walking	d4500 – Walking short distances *Related: b735 Muscle Tone, b755 Balance*
	Enable active exercise (exercise programme)	5 - Self-Care	d570 – Looking after health	d5702 – Maintaining ones health *Related: b735 Muscle Tone, b710 Mobility of joints, b760 Control of voluntary movement functions*
	Enable hygiene/Self Care	5 - Self-Care	d520 – Caring for body parts	d5208 – Caring for body parts, specified *Related: b810 Protection of skin, b820 Repair of skin*
	Enable splint application	5 - Self-Care	d520 – Caring for body parts	d5208 – Caring for body parts, specified *Related: b735 Muscle Tone, b710 Mobility of joints*
	Enable postural management	5 - Self-Care	d520 – Caring for body parts	d5208 – Caring for body parts, specified *Related: b735 Muscle Tone, b710 Mobility of joints*
Other				
	Improve cosmesis of arm		Does not fit in an ICF category	

Key: ***Related : Body function domains related to achievement of activity goals***

Goals in the cohort study

Goals set during the cohort study were in the domains of body function and activity, with one goal set (improving arm cosmesis) which did not fit the framework of the ICF. The activity goal categories broadly matched those covered by the ArmA and were in mobility and self care ICF chapters. This further supports the use of the ArmA in groups undergoing focal spasticity management intervention.

Some differences exist in the goals set in the current cohort study and those set in the study by Turner-Stokes and colleagues (Turner-Stokes et al. 2010). For example, the ArmA does not address community and social activities. However, both community and social activities are primarily participation issues rather than activity as defined by the ICF. The difference between the goals in the studies also possibly reflects the overall high level of impairment and dependency of the patient group studied in this thesis, compared with the groups studied in some other trials (Brashear et al. 2002; Turner-Stokes et al. 2010).

The ArmA has been developed to address specific issues seen in focal spasticity intervention where active and passive improvements may be made. Improvements in passive function have particular importance for patients with more severe neurological disability who will usually not improve in active function, but for whom ease of care is particularly important.

8.2.2 Strengths and challenges of the findings

The following section will address the strengths and challenges of the findings in the following three areas; application of the ArmA measure; the relationship of passive function to active function and physical therapy contribution to functional improvements in focal spasticity intervention.

8.2.2.1 Application of the ArmA

Most clinicians in rehabilitation practice will be familiar with the use and application of outcome measures. While in some teams outcome measurement may be infrequently used, current standards of practice such as codes of professional practice (Chartered

Society of Physiotherapy 2008) and the Royal College of Physicians – *Spasticity in adults: management using botulinum toxin*, (Royal College of Physicians et al. 2009) emphasise the use of outcome measures. The Darzi review, *High Care Quality for All*, places the use of clinical outcome measures firmly in the context of current policy direction for health care in general (Department of Health 2008). Outcome measures are therefore central to rehabilitation practice, physical therapy intervention and decision-making about ongoing management (Playford et al. 2009; Wissel et al. 2009). Measuring function in clinical practice has been emphasised as a priority in spasticity management and other areas of rehabilitation practice because functional improvement is often a main goal (Royal College of Physicians et al. 2009; Wissel et al. 2009).

Sheean was one of the first authors to refer to the terms active and passive function in the literature and used the terms to describe the different types of improvement seen clinically following spasticity intervention (Sheean 2001). In the same paper Sheean posed the question *"why is it difficult to show a functional benefit"* following BTX intervention for spasticity? In suggesting possible answers to this question, he proposed that it may be difficult to show functional benefit because of a lack of appropriate measures which are sensitive to change. The cohort study has demonstrated that it is possible to show passive function change following spasticity intervention using the ArmA.

8.2.2.2 Relationship of passive function to active function

The cohort study has provided further confirmation that, as expected, improvements in passive function do not automatically lead to improvements in active function. Passive function improvements relate more directly to reduction in spasticity, which allows improved passive movement of the affected upper limb, making care tasks easier. Improvements in passive function may be further augmented by physical interventions, which in the case of casting and splinting maintain or increase joint range through mechanical extension and longer duration 'stretch' (Lannin and Herbert 2003).

Improvements in passive function therefore occur largely because of an increase in the available range of movement produced through spasticity reduction and increases in true range of movement due to stretch following physical interventions. The changes

can directly influence passive function by increasing the ease of extending the affected limb for care tasks. For example, cleaning the axilla of the affected upper limb will be easier following injection of BTX, combined with positioning of the upper limb, because it is then much easier to passively abduct the arm and clean the armpit.

Improvements in active function rely on the mobility of the joint or limb and the ability to recruit movement and muscle strength in a selective manner allowing purposeful activity. If joint range of movement is lost due to muscle being in a shortened position for a prolonged period of time (e.g. as a result of spasticity) then movement and function will be inhibited. However, to improve active function, selective control by the patient will also need to produce purposeful movement.

Selective control will not necessarily follow improved range of movement because it relies on potential for recovery dependent on the extent of the initial injury and the plasticity of the central nervous system. The purpose of BTX is the weakening of muscle. Therefore, direct improvement in active movement and therefore function is not possible. Active function in the right circumstances may be progressed and developed by task practice, but the ability to engage in this type of practice is needed (Dettmers et al. 2005; Uswatte and Taub 2005). Active function improvement will only be possible if some degree of control is present or can be re-learned to allow practice and secondary problems such as spasticity or contracture are prevented or reversed. Active function improvement may be particularly relevant in BTX intervention provided soon after initial neurological insult in selected patients (Cousins et al. 2010; Hesse et al. 2011), although this requires further evaluation.

8.2.2.3 The contribution of physical therapy to functional improvements

In addition to Modified CIMT (MCIMT) and task practice, the PT intervention categories in this thesis consisted of splinting (non-circumferential and removable orthotics), serial casting (circumferential removable and non-removable orthotics), positioning of the upper limb, and functional electrical stimulation (FES). However interventions, which have an initial impact on passive function, deserve further evaluation in the context of management of spasticity in the hemiparetic upper limb (Bergfeldt et al. 2006). Of particular interest for further evaluation are splinting (Lannin

and Herbert 2003; Lannin et al. 2007), serial casting and passive stretch (Hill 1994; Moseley et al. 2008). Trials to evaluate these interventions should include a self-report measure such as the ArmA to assess passive function outcome in the context of real-life.

Further exploration of the combination of BTX with 'task practice' such as Constraint Induced Movement Therapy (CIMT) (Taub et al. 1993; Dettmers et al. 2005) would be valuable. However, 'task practice' is only appropriate in the small number of individuals undergoing BTX intervention for management of spasticity who have potential for improvement in active function. There is therefore a stronger clinical need to explore application of task practice intervention in the total brain injury population rather than confining evaluation just to those undergoing spasticity management who will be a small minority. Future development of the ArmA active function sub-scale could involve evaluation and validation in these wider applications.

8.3 Future development of the ArmA measure

The following section focuses on developments in ArmA evaluation and its application in addressing some of the research questions posed. Future research and development broadly fits into two areas; further evaluation of the properties of the ArmA measure and further evaluation of upper limb interventions using the ArmA measure.

8.3.1 Further evaluation of the ArmA measure

Future research evaluating the ArmA measure should focus on three areas; 1) confirming changes made to the active and passive function sub-scales, 2) further evaluation of the measurement properties of the active and passive sub-scales using IRT methods and 3) evaluation of interval scaling and longitudinal construct validity of the ArmA using IRT and CTT methods.

1) It is necessary to consider modifications to the active and passive function sub-scales (for further details see Section 7.6.1.2; page 243). Further evaluation of the passive function and active function sub-scales will be ongoing using data from routine practice in our setting resulting in enlarging the current data set and allowing initial exploration of the changes made (Harrow Ethics committee 04/Q0405/81).

2) Mokken analysis, should be re-applied using the double monotonicity model once a larger cohort of participants has been recruited to confirm the findings of the preliminary evaluation of the passive function sub-scale in this thesis. The active function sub-scale requires evaluation of its dimensionality and measurement scaling properties by applying the ArmA in a more able group of participants not undergoing spasticity intervention. Undertaking evaluation in patients not undergoing spasticity intervention will enable this work to be completed in a timely manner, rather than the extremely prolonged process if evaluated in those undergoing spasticity intervention.

3) Following initial analysis of ordinal scaling using the double monotonicity model, the interval scaling properties of the ArmA can be evaluated using the Rasch method. Application of Rasch analysis, if applied in longitudinal data (e.g. baseline, 8 weeks and 16 weeks as undertaken in this thesis), will in due course allow the evaluation of longitudinal construct validity and identification of minimal important change in the active and passive sub-scales. Using latent variable methods, evaluation of the consistency of the measurement properties across the three time points can also be undertaken using confirmatory factor analysis.

8.3.2 Further evaluation of functional improvement following spasticity management

Bakheit and colleagues have identified changes in passive function maintained up to 16 weeks (Bakheit et al. 2000; Bakheit et al. 2001). Francis and colleagues also suggested that in some patients, maximal passive function improvement, may occur sometime after maximal improvement in spasticity following BTX intervention (Francis et al. 2004). These issues require further evaluation in due course.

Passive function

For improvement in passive function two areas of future research have been identified:

- Which physical therapy interventions (see Appendix 18) and dosages are most effective in producing and maintaining passive function improvement?

- For what time period is it possible to maintain these improvements with physical therapy intervention?

Active function

For active function it is important to more fully understand which patients are likely to achieve active function improvement following BTX intervention and what combination of interventions are effective in producing that change. For patients with active function goals the following research areas have been identified:

- Which patients will benefit from spasticity intervention as a contribution towards active function improvement?
- Which interventions to improve task performance are most effective in producing maximum improvement in active functional outcome?

Before undertaking, a further investigation of these areas it will be necessary to develop an upper limb treatment-recording schedule to record, in a systematic manner, the physical therapy interventions provided to patients. Donaldson has undertaken similar work in an investigation of conventional physiotherapy with functional strength training for rehabilitation of the upper limb after stroke (Hunter et al. 2006; Donaldson 2007; Donaldson et al. 2009) and related work has also been undertaken by De Wit (De Wit et al. 2007) (also see Section 7.6.2; page 254).

8.4 Conclusions

This thesis has made three main contributions to current knowledge.

- Firstly, the ArmA, a new measure of active and passive function, with potential for both clinical and research use, was developed and underwent preliminary psychometric testing including evaluation of ordinal scaling.
- Secondly, the complexities of measuring active and passive function as latent variables have been explored with reference to psychometric methods.
- Thirdly, further evidence was generated indicating that improvements in passive function occur following BTX and PT interventions by 8 weeks and that these are maintained at 16 weeks despite a re-increase in spasticity as measured by the Modified Ashworth Scale (MAS). Further exploration of these preliminary findings will inform future developments in clinical practice.

The overall hypotheses (see Chapter 1; page 59) for the thesis were accepted, as follows:

1. Items have been successfully identified from literature and clinical practice sources leading to acceptance of hypothesis 1.
2. Reduction of these items using Delphi and wider consultation processes with patients, carers and clinicians has been successfully undertaken resulting in a draft measure and leading to acceptance of hypothesis 2.
3. Preliminary evidence for the passive function items, that the sub-scale scores form a measurement system, capable (following further evaluation) of providing a single summary statistic has been provided. However, hypothesis 3 cannot be fully accepted on the current evidence and requires further evaluation particularly of the active function sub-scale.
4. In the cohort study, the ArmA has been used to demonstrate improvement in passive function following treatment of upper limb spasticity using BTX and PT intervention leading to acceptance of hypothesis 4.

In conclusion, the ability to measure function following focal upper limb spasticity intervention has been expanded following the development and preliminary testing of the ArmA. The measurement properties of the ArmA now require further evaluation, with particular emphasis on a more complete evaluation of the active function sub-scale and establishing interval level scaling in both sub-scales.

Further understanding of the effects of spasticity management on function has also been developed. Physical interventions may both contribute to the initial achievement of functional goals and maintain these improvements in the longer term. The exact duration and process of maintenance require further investigation. This work has raised questions about the possible mechanisms of functional change, and raises new questions for future evaluation and research.

References

Aleamoni, L. M. (1976). The relation of sample size to the number of variables in using factor analysis techniques. Educational and Psychological Measurement. **36**: 879-883.

Alon, G., Dar, A., Katz-Behiri, D., Weingarden, H. and Nathan, R. (1998). Efficacy of a hybrid upper limb neuromuscular electrical stimulation system in lessening selected impairments and dysfunctions consequent to cerebral damage. Journal of Neurological Rehabilitation. **12**(2): 73-79.

Alon, G., Sunnerhagen, K. S., Geurts, A. C. H. and Ohry, A. (2003). A home-based, self-administered stimulation program to improve selected hand functions of chronic stroke. NeuroRehabilitation. **18**(3): 215-225.

Altman, D. G. (1991). Practical statistics for medical research. London, Chapman and Hall.

Andersen, E. B. (1977). Sufficient statistics and latent trait models. Psychometrika. **42**(1): 69-81.

Apgar, V. (1953). A proposal for a new method of evaluation of the newborn infant. Current researches in anaesthesia & analgesia . **32**(4): 260–267.

Ashford, S., Slade, M., Malaparade, F. and Turner-Stokes, L. (2008). Evaluation of functional outcome measures for the hemiparetic upper limb: a systematic review Journal of Rehabilitation Medicine. **40**(9): 787-795.

Ashford, S. and Turner-Stokes, L. (2006). Goal attainment for spasticity management using botulinum toxin. Physiotherapy Research International. **11**(1): 24-34.

Ashford, S. and Turner-Stokes, L. (2008). Management of shoulder and proximal upper limb spasticity using botulinum toxin and concurrent therapy interventions: A preliminary analysis of goals and outcomes. Disability and Rehabilitation. **31**: 220-226.

Australian Institute of Health and Welfare (2003). ICF Australian user guide version 1.0, Australian Institute of Health and Welfare.

Bakheit, A. M., Pittock, S., Moore, A. P., Wurker, M., Otto, S., Erbguth, F. and Coxon, L. (2001). A randomised, double-blind, placebo-controlled study of the efficacy and safety of botulinum toxin type A in upper limb spasticity in patients with stroke. European Journal of Neurology. **8**(6): 559-565.

Bakheit, A. M., Thilmann, A. F., Ward, A. B., Poewe, W., Wissel, J., Muller, J., Benecke, R., Collin, C., Muller, F., Ward, C. D. and Neumann, C. (2000). A randomised, double-blind, placebo-controlled, dose-ranging study to compare the efficacy and safety of three doses of botulinum toxin type A (Dysport) with placebo in upper limb spasticity after stroke. Stroke. **31**(10): 2402-2406.

Bakheit, A. M. O. (2004a). Optimising the methods of evaluation of the effectiveness of botulinum toxin treatment of post-stroke muscle spasticity. Journal of Neurology, Neurosurgery and Psychiatry. 665-666.

Bakheit, A. M. O., Fedorova, N. V., Skoromets, A. A., Timerbaeva, S. L., Bhakta, B. B. and Coxon, L. (2004b). The beneficial antispasticity effect of botulinum toxin type A is maintained after repeated treatment cycles. Journal of Neurology, Neurosurgery & Psychiatry. **75**(11): 1558-1561.

Barnes, M. (2003). Botulinum toxin -- mechanisms of action and clinical use in spasticity. Journal of Rehabilitation Medicine. **9**(25): 1650-1677.

Barnes, M., Schnitzler, A., Medeiros, L., Aguilar, M., Lehnert-Batar, A. and Minnasch, P. (2010). Efficacy and safety of NT 201 for upper limb spasticity of various etiologies - a randomised parallel-group study. Acta Neurologica Scandinavica. **122**: 295-302.

Barreca, S. R., Stratford, P. W., Lambert, C. L., Masters, L. M. and Streiner, D. L. (2005). Test-Retest Reliability, Validity, and Sensitivity of the Chedoke Arm and Hand Activity Inventory: A New Measure of Upper-Limb Function for Survivors of Stroke

Archives of Physical Medicine and Rehabilitation. **86**(8): 1616-1922.

Barrett, P. and Kline, P. (1981). A comparison between Rasch analysis and factor analysis of items in the EPQ. Journal of Personality and Group Behaviour. **1**: 1-21.

Beaton, D. E., Katz, J. N., Fossel, A. H., Wright, J. G., Tarasuk, V. and Bombardier, C. (2001). Measuring the whole or the parts? Validity, reliability and responsiveness of the Disabilities of the Arm Shoulder and Hand outcome measure in different regions of the upper limb. Journal of Hand Therapy. **14**(2): 128-146.

Beaton, D. E., Wright, J. and Katz, D. I. (2005). Development of the quick DASH: Comparison of three item-reduction approaches. Journal of Bone and Joint Surgery. **87**(5): 1038-1046.

Bedard, M., Molloy, D. W., Squire, L., Dubois, S., Lever, J. A. and O'Donnell, M. (2001). The Zarit Burden Interview: a new short version and screening questionnaire. The Gerontologist. **41**: 249-258.

Bell, K. R. and Williams, F. (2003). Use of botulinum toxin type A and type B for spasticity in upper and lower limbs. Physical Medicine & Rehabilitation Clinics of North America. **14**(4): 821-835.

Bender, D. E. and McKenna, K. (1994). The focus group as a tool for health research: issues in design and analysis. Health Transition Review. **4**(1): 63-76.

Bergfeldt, U., Borg, K., Kullander, K. and Julin, P. (2006). Focal spasticity therapy with botulinum toxin: Effects on function, activities of daily living and pain in 100 adult patients. Journal of Rehabilitation Medicine. **38**: 166-171.

Berglund, K. and Fugl-Meyer, A. R. (1986). Upper extremity function in hemiplegia. A cross-validation study of two assessment methods. Scandinavian Journal of Rehabilitation Medicine. **18**(4): 155-157.

Bhakta, B. B. (2000b). Management of spasticity in stroke. British Medical Bulletin. **56**(2): 476-485.

Bhakta, B. B., Cozens, J. A., Bamford, J. M. and Chamberlain, M. A. (1996). Use of botulinum toxin in stroke patients with severe upper limb spasticity. Journal of Neurology, Neurosurgery & Psychiatry. **61**(1): 30-35.

Bhakta, B. B., Cozens, J. A., Chamberlain, M. A. and Bamford, J. M. (2000a). Impact of botulinum toxin type A on disability and carer burden due to arm spasticity after stroke: a randomised double-blind placebo-controlled trial. Journal of Neurology, Neurosurgery and Psychiatry. **69**(2): 217-221.

Bhakta, B. B., O'Connor, R. J. and Cozens, J. A. (2008). Associated reactions after stroke: a randomised controlled trial of the effect of botulinum toxin type A. Journal Rehabilitation Medicine. **40**: 36-41.

Biemans, M. A. J. E., Dekker, J. and van der Woude, L. H. V. (2001). The internal consistency and validity of the Self-assessment Parkinson's Disease Disability Scale. Clinical Rehabilitation. **15**: 221-228.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F. M. and Novack, T. A. Statistical theories of mental test scores. Reading, Addison-Wesley.

Blake, P. F. and Fritz, V. U. (1996). Functional outcome of the upper limb after stroke. South African Journal of Physiotherapy. **52**(2): 40-42.

Bland, J. M. and Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement Lancet. **i**: 307-310.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika. **37**: 29-51.

Boiteau, M., Malouin, F. and Richards, C. L. (1997). The painful hemiplegic shoulder: effects of intra-articular triamcinolone acetonide. American journal of Physical Medicine & Rehabilitation / Association of Academic Physiatrists. **76**(1): 43-48.

Borsboom, D. (2005). Measuring the mind; Conceptual issues in contemporary psychometrics. Cambridge, Cambridge University Press.

Borsboom, D. (2006). The attack of the psychometricians. Psychometrika. **71**(3): 425-440.

Bot, S. D., Terwee, C. B., van der Windt, D. A., Bouter Lex, M. and deVet, H. C. W. (2004). Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature. Annals of the Rheumatic Diseases **63**: 335-341.

Brashear, A., Gordon, M. F., Elovic, E., Kassicieh, V. D., Marciniak, C., Do, M., Lee, C., Jenkins, S. and Turkel, C. (2002a). Intramuscular injection of botulinum toxin for the treatment of wrist and finger spasticity after a stroke. New England Journal of Medicine. **347**(6): 395-400.

Brashear, A., McAfee, A. L., Kuhn, E. R. and Fyffe, J. (2004). Botulinum toxin type B in upper-limb post-stroke spasticity: A double-blind, placebo-controlled trial. Archives of Physical Medicine & Rehabilitation. **85**(5): 705-709.

Brashear, A., Zafonte, R., Corcoran, M., Galvez-Jimenez, N., Gracies, J. M., Gordon, M. F., McAfee, A., Ruffing, K., Thompson, B., Williams, M., Lee, C. H. and Turkel, C. (2002b). Inter- and intrarater reliability of the Ashworth Scale and the Disability Assessment Scale in patients with upper-limb post-stroke spasticity. Archives of Physical Medicine & Rehabilitation. **83**(10): 1349-54.

Brin, M. F. (1997). Dosing, administration, and a treatment algorithm for use of botulinum toxin A for adult-onset spasticity. Spasticity Study Group. Muscle & Nerve (Supplement). **6**: S208-S220.

Broeks, J. G., Lankhorst, G. J., Rumping, K. and Prevo, A. J. (1999). The long-term outcome of arm function after stroke: results of a follow-up study. Disability and Rehabilitation. **21**(8): 357-364.

Burke, W., Wesolowski, M. and Guth, M. (1988). Comprehensive head injury rehabilitation: an outcome evaluation. Brain Injury. **2**: 313-322.

Burns, S. P., Rivara, F. P., Johansen, J. M. and Thompson, D. C. (2003). Rehabilitation of traumatic injuries. Use of the Delphi method to identify topics for evidence-based review. American Journal of Physical Medicine & Rehabilitation. **82**(5): 410-414.

Butler, L. M., London, S. J., Yu, M. C., Tseng, M., Koh, W.-P. and Lee, H.-P. (2006). On the usage of Principle Components Analysis and Multiple Testing. American Journal of Respiratory and Critical Care Medicine. **173**: 574-575.

Campbell, N. R. (1920). Physics, the elements. Cambridge, Cambridge University Press.

Cano, S., Warner, T. T., Thompson, A. J., Bhatia, K. P., Fitzparick, R. and Hobart, J. C. (2008). The Cervical Dystonia Impact Profile (CDIP-58): Can a Rasch developed patient reported outcome measure satisfy traditional psychometric criteria? Health and Quality of Life Outcomes. **6**(58).

Carda, S. and Molteni, F. (2005). Taping versus electrical stimulation after botulinum toxin type A injection for wrist and finger spasticity. A case-control study. Clinical Rehabilitation. **19**: 621-626.

Carroll, D. (1965). A quantitative test of upper limb extremity function. Journal of Chronic Diseases. **18**: 479-491.

Chang, C. L., Munin, M. C., Skidmore, E. R., Niyonkuru, C., Huber, L. M. and Weber, D. J. (2009). Effect of baseline spastic hemiparesis on recovery of upper-limb function following botulinum toxin type a injections and post injection therapy. Archives of Physical Medicine and Rehabilitation. **90**: 1462-1468.

Chartered Society of Physiotherapy (2008). Chartered Society of Physiotherapy. Core Standards of Physiotherapy Practice. London, UK., Chartered Society of Physiotherapy

Childers, M. K., Brashear, A., Jozefczyk, P., Reding, M., Alexander, D., Good, D., Walcott, J. M., Jenkins, S. W., Turkel, C. and Molloy, P. T. (2004). Dose-dependent response to intramuscular botulinum toxin type A for upper-limb spasticity in patients after a stroke. Archives of Physical Medicine and Rehabilitation. **85**(7): 1063-1069.

Cieza, A., Brockow, T., Ewert, T., Amman, E., Kollerits, B., Chatterji, S., Ustun, B. and Stucki, G. (2002). Linking health-status measurements to the international classification of functioning, disability and health. Journal of Rehabilitation Medicine. **34**: 205-210.

Cohen, J. (1988). Statistical power analysis for the behavioural sciences. Hillsdale, NJ, Lawrence Erlbaum.

Comfrey, A. L. and Lee, H. B. (1992). A first course in Factor analysis Hillsdale USA, Lawrence Erlbaum Associates.

Coombs, C. H. (1950). Psychological scaling without limit. Psychological Review. **57**: 145-158.

Cousins, E., Ward, A. B., Roffe, C., Rimington, L. and Pandyan, A. D. (2010). Does low-dose botulinum toxin help the recovery of arm function when given early after stroke? A phase II randomized controlled pilot study to estimate effect size. Clinical Rehabilitation. **24**: 501-513.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika. **16**: 297-334.

Cronbach, L. J. and Meehl, P. E. (1955). Construct validity in psychological tests. First published in Psychological Bulletin. **52**: 281-302.

Cusick, A., Vasquez, M., Knowles, L. and Wallen, M. (2005). Effect of rater training on reliability of Melbourne Assessment of Unilateral Upper Limb Function scores. Developmental Medicine & Child Neurology. **47**(1): 39-45.

Darlington, R. B. (2010). Factor Analysis. http://www.psych.cornell.edu/darlington/factor.htm.

Davis, A. M., Beaton, D. E., Hudak, P. and Amadio, P. (1999). Measuring disability of the upper extremity: a rationale supporting the use of a regional outcome measure. Journal of Hand Therapy. **12**: 269-274.

Davis, E. C. and Barnes, M. (2000). Botulinum toxin and spasticity. Journal of Neurology, Neurosurgery and Psychiatry. **69**: 143-149.

de Koning, E., Sijtsma, K. and Hamers, J. H. M. (2002). Comparison of four IRT models when analysing two tests for inductive reasoning. Applied Psychological Measurement. **26**(3): 302-320.

De Wit, L., Kamsteegt, H., Yadav, B., Verheyden, G., Feys, H. and De Weerdt, W. (2007). Defining the content of individual physiotherapy and occupational therapy sessions for stroke patients in an inpatient rehabilitation setting. Development, validation and inter-rater reliability of a scoring list. Clinical Rehabilitation. **21**: 450-459.

Deane, K. H. O., Ellis-Hill, C., Dekker, K., Davies, P. and Clarke, C. E. (2003). A Delphi survey of best practice occupational therapy for Parkinson's disease in the United Kingdom. British Journal of Occupational Therapy. **66**: 247-254.

Deary, I. J., Wilson, J. A., Carding, P. N., MacKenzie, K. and Watson, R. (2010). From dysphonia to dysphoria: Mokken scaling shows a strong, reliable hierarchy of voice symptoms in the voice symptom scale questionnaire. Journal of Psychosomatic Research. **68**: 67-71.

DeCoster, J. (1998). Overview of Factor Analysis. Downloaded 22/12/2010 from http://www.stat-help.com.

DeJong, A. and Molenaar, I. W. (1987). An application of Mokken's model for stochastic, cumulative scaling in psychiatric research. Journal of Psychiatric Research. **21**(2): 137-149.

DeJong, G., Horn, S. D., Conroy, B., Nichols, D. and Healton, E. B. (2005). Opening the black box of post stroke rehabilitation: Stroke rehabilitation patients, processes and outcomes. Archives of Physical Medicine and Rehabilitation. **86**(Supplement 2): S1-S7.

DeJong, G., Horn, S. D., Gassaway, J. A., Slavin, M. D. and Dijkers, M. P. (2004). Toward a taxonomy of rehabilitation interventions: Using an inductive approach to examine the "black box" of rehabilitation. Archives of Physical Medicine and Rehabilitation. **85**: 678-686.

Department of Health (2008). Department of Health High Quality Care for All. NHS Next Stages Review Final Report. (The Darzi Report). London, Available from: http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationspolicyandGuidance/DH-085825 (Accessed 15th March 2010).

Department of Health (2009). Public and patient experience and engagement, Department of Health. London. Available from: http://www.dh.gov.uk/en/Managingyourorganisation/PatientAndPublicinvolvement/DH_098642 (Accessed 14th December 2009).

DeSouza, L. H., Langton-Hewer, R. and Miller, S. (1980). Assessment of recovery of arm control in hemiplegic stroke patients. Arm function test. International Rehabilitation Medicine. **2**: 3-9.

Desrosiers, J., Bravo, G., Hebert, R., Dutil, E. and Mercier, L. (1994). Validation of the box and block test as a measure of dexterity of elderly people: Reliability,

Validity and Norms Studies. Archives of Physical Medicine and Rehabilitation. **75**: 751-755.

Desrosiers, J., Hebert, R., Bravo, G. and Dutil, E. (1995). The Purdue pegboard test: normative data for people aged 60 and over. Disability and Rehabilitation. **17**: 217-224.

Dettmers, C., Teske, U., Hamzei, F., Uswatte, G., Taub, E. and Weiller, C. (2005). Distributed form of constraint-induced movement therapy improves functional outcome and quality of life after stroke. Archives of Physical Medicine and Rehabilitation. **86**(2): 204-209.

DeVellis, R. F. (2006). Classical test theory. Medical Care. **44**(11 Supplement 3).

Dickersin, K., Scherer, R. and Lefebvre, C. (1994). Systematic Reviews: Identifying relevant studies for systematic reviews. British Medical Journal. **309**(12 November): 1286-1291.

Dickson, H. G. and Kohler, F. (1996). The multi-dimensionality of the FIM motor items precludes an interval scaling using Rasch analysis.[comment]. Scandinavian Journal of Rehabilitation Medicine. **28**(3): 159-162.

Donabedian, A. (1980). The Definition of Quality and Approaches to Its Assessment Ann Arbor, MI: Health Administration Press.

Donaldson, C. (2007). An investigation of conventional physiotherapy and functional strength training for rehabilitation of the upper limb after stroke (dissertation). Rehabilitation. London, University of London, St George's**:** 314.

Donaldson, C., Tallis, R. and Pomeroy, V. M. (2009). A treatment schedule of conventional physical therapy provided to enhanced upper limb sensorimotor recovery after stroke: Expert criterion validity and intra-rater reliability. Physiotherapy. **95**: 110-119.

Dressler, D., Saberi, F. A. and Barbosa, E. R. (2005). Botulinum toxin - Mechanisms of action. Arquivos de Neuro-Psiquiatria **63**(1): 180-185.

Edgeworth, F. Y. (1888). The statistics of examinations. Journal of the Royal Statistical Society. **51**: 598-635.

Elia, A. E., Filippini, G., Calandrella, D. and Albanese, A. (2009). Botulinum neurotoxins for post-stroke spasticity in adults: a systematic review. Movement Disorders. **24**(6): 801-812.

Elovic, E., A, B., Kaelin, D., Liu, J., Millis, S. R., Barron, R. and C, T. (2008). Repeated treatments with botulinum toxin A produce sustained decreases in the

limitations associated with focal upper-limb post stroke spasticity for caregivers and patients. Archives of Physical Medicine and Rehabilitation. **89**: 799-806.

Embretson, S. E. (1996). The new rules of measurement. Psychological Assessment. **8**: 341-349.

Embretson, S. E. and Reise, S. P. (2000). Item Response Theory for Psychologists. Mahwah, New Jersey, Lawrence Erlbaum Associates.

Fayers, P. M. and Machin, D. (2007). Quality of life: the assessment, analysis, and interpretation of patient-reported outcomes. Chichester, John Wiley & Sons Ltd.

Feinstein, A. R. (1983). An additional basic science for clinical medicine: IV. The development of clinimetrics. Annals of Internal Medicine. **99**(6): 843-848.

Feinstein, A. R. (1987). Clinimetric perspectives. Journal of Chronic Diseases. **40**(6): 635-640.

Finger, M., Cieza, A., Stoll, J., Stucki, G. and Huber, E. O. (2006). Identification of intervention categories for physical therapy, based on the international classification of functioning, disability and health: A Delphi exercise. Physical Therapy. **86**(9): 1203-1220.

Fitzpatrick, R., Davey, C., Buxton, M. J. and Jones, D. (1998). Evaluating patient-based outcome measures for use in clinical trials. Health Technologies Assessment. **2**(14).

Fleiss, J. L. (1981). Statistical methods for rates and proportions. New York, John Wiley & Sons.

Forkmann, T., Boecker, M., Norra, C., Eberle, N., Kircher, T., Schauerte, P., Mischke, K., Westhofen, M. and Witz, M. (2009). Development of an item bank for the assessment of depression in persons with mental illness and physical diseases using Rasch analysis. Rehabilitation Psychology. **54**(2): 186-197.

Forkmann, T., Boecker, M., Wirtz, M., Eberle, N., Westhofen, M., Schauerte, P., Mischke, K., Kircher, T., Gauggel, S. and Norra, C. (2009b). Development and validation of the Rasch based depression screening (DESC) using Rasch analysis and structural equation modelling. Journal of Behavioural Therapy and Experimental Psychiatry. **40**: 469-478.

Francis, H. P., Wade, D. T., Turner-Stokes, L., Kingswell, R. S., Dott, C. S. and Coxon, E. A. (2004). Does reducing spasticity translate into functional benefit? An exploratory meta-analysis. Journal of Neurology, Neurosurgery & Psychiatry. **75**(11): 1547-1551.

Gassaway, J. A., Horn, S. D., DeJong, G., Smout, R. J., Clark, C. and James, R. (2005). Applying the clinical practice improvement approach to stroke rehabilitation: Methods used and baseline results. Archives of Physical Medicine and Rehabilitation. **86**(Supplement 2): S16-S33.

Geyh, S., Cieza, A., Schouten, J., Dickson, H., Frommelt, P., Omar, Z., Ring, H., Kostanjsek, N. and Stucki, G. (2004). ICF core data sets for stroke. Journal of Rehabilitation Medicine. **44 (Supplement)**: 135-141.

Gillespie, M., Tenvergert, E. M. and Kingma, J. (1987). Using Mokken scale analysis to develop unidimensional scales. Quality and Quantity. **21**: 393-408.

Giordano, A., Pucci, E., Naldi, P., Mendozzi, L., Milanese, C., Tronci, F., Leone, M., Mascoli, N., La Mantia, L., Giuliani, G. and Solari, A. (2009). Responsiveness of patient reported outcome measures in multiple sclerosis relapses: the REMS study. Journal of Neurology, Neurosurgery and Psychiatry **80**: 1023-1028.

Giovannelli, M., Borriello, G., Castri, P., Prosperi, L. and Pozzilli, C. (2007). Early physiotherapy after injection of botulinum toxin increases the beneficial effects on spasticity in patients with multiple sclerosis. Clinical Rehabilitation. **21**: 331-337.

Gompertz, P., Pound, C. and Ebrahim, S. (1994). A postal version of the Barthel index. Clinical Rehabilitation. **8**(3): 233-239.

Gorsuch, R. L. (1983). Factor analysis (second edition). Hillsdale, USA, Lawrence Erlbaum Associates.

Granger, C. V., Cotter, A. C., Hamilton, B. B. and Fiedler, R. C. (1993). Functional assessment scales: A study of persons after stroke. Archives of Physical Medicine & Rehabilitation. **74**(2): 133-138.

Green, B. F. (1954). Attitude measurement. In. Handbook of social psychology. Reading, Addison-Wesley. **1:** 335-369.

Greenhalgh, J., Long, A. F. and Flynn, R. (2005). The use of patient reported outcome measures in routine clinical practice: lack of impact or lack of theory? Social Science and Medicine. **60**(4): 833-843.

Guadagnoli, E. and Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. Psychological Bulletin. **103**: 265-275.

Guttman, L. (1950). The basis of scalogram analysis. In Stoufer, S. A., Guttman, L., Suchman, E. A. et al. Studies in social psychology in World War II: Volume 1.

Measurement and prediction. Princeton, USA, Princeton University Press**:** 60-90.

Guyatt, G. H. (1987). Measuring change over time: Assessing the usefulness of evaluative instruments. Journal of Chronic Diseases. **40**: 171-178.

Hagquist, C., Bruce, M. and Gustavsson, J. P. (2009). Using the Rasch model in nursing research: An introduction and illustrative example. International Journal of Nursing Studies. **46**: 380-393.

Haig, B. D. and Borsboom, D. (2008). On the conceptual foundations of psychological measurement. Measurement. **6**: 1-6.

Hambleton, P. and Moore, A. P. (1995). Botulinum neurotoxins: origin, structure, molecular actions and antibodies. In Moore, A. P. In: Handbook of botulinum toxin treatment. Oxford, Blackwell Science**:** 16-27.

Hanley, B., Bradburn, J., Barnes, M., Evans, C., Goodare, H., Kelson, M., Kent, A., Oliver, S., Thomas, S. and Wallcraft, J. (2004). Involving the public in NHS, public health and social care research: briefing notes for researchers. . INVOLVE. Available from: http://www.invo.org.uk/posttypepublications/briefing-note-for-researchers (accessed 8th October 2009).

Hart, T. and Evans, J. (2006). Self-regulation and goal theories in brain injury rehabilitation. Journal of Head Trauma Rehabilitation. **21**(2): 142-155.

Hatcher, L. (1994). A step-by-step approach to using the SAS system for factor analysis and structural equation modelling. Cary, N C, SAS Institute, Inc.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. Applied Psychological Measurement. **9**: 139-164.

Hays, R. D., Brown, J., Brown, L. U., Spritzer, K. L. and Crall, J. J. (2006). Classical test theory and item response theory analyses of multi-item scales assessing parents' perceptions of their children's dental care. Medical Care. **44**(11 Supplement 3): S60-S68.

Hays, R. D., Morales, L. S. and Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. Medical Care. **38**(9 Supplement): 1128-1142.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. Measurement and Evaluation in Counselling and Development. **34**: 177-189.

Hesse, S., Mach, H., Frohlich, S., Behrend, S., Werner, C. and Melzer, I. (2011). An early botulinum toxin A treatment in sub-acute stroke patients may prevent a disabling finger flexor stiffness six months later: a randomized controlled trial. Clinical Rehabilitation. 0269215511421355, first published on October 4, 2011.

Hesse, S., Reiter, F., Konrad, M. and Jahnke, M. T. (1998). Botulinum toxin type A and short-term electrical stimulation in the treatment of upper limb flexor spasticity after stroke: a randomised, double-blind placebo-controlled study. Clinical Rehabilitation. **12**: 381-388.

Hicks, C. M. (1999). Research methods for clinical therapists; Applied project design and analysis, Churchill Livingstone; London.

Higgins, J., Mayo, N. E., Desrosiers, J., Salbach, N. M. and Sara Ahmed (2005). Upper-limb function and recovery in the acute phase post-stroke. Journal of Rehabilitation Research & Development. **42,**(1): 65-76.

Hill, J. (1994). The effects of casting on upper extremity motor disorders after brain injury. American Journal of Occupational Therapy. **48**(3): 219-224.

Hobart, J. C. and Cano, S. (2009). Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. Health Technologies Assessment. **13**(12): 1-200.

Hobart, J. C., Cano, S. and Thompson, A. J. (2010). Effect sizes can be misleading: Is it time to change the way we measure change. Journal of Neurology Neurosurgery and Psychiatry. **81**: 1044-1048.

Hobart, J. C., Cano, S., Zajicek, J. P. and Thompson, A. J. (2007). Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. Lancet Neurology. **6**: 1094-1105.

Hobart, J. C., Lamping, D. L. and Freeman, J. A. (2001). Evidence-based measurement: which disability scale for neurological rehabilitation? Neurology. **57**: 639-644.

Hobart, J. C., Lamping, D. L. and Thompson, A. J. (1996). Evaluating neurological outcome measures: the bare essentials. Journal of Neurology, Neurosurgery and Psychiatry. **60**: 127-130.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika. **32**: 179-185.

Hudak, P., Amadio, P. and Bombardier, C. (1996). Development of an upper extremity outcome measure: the DASH (disabilities of the arm, shoulder and hand). American Journal of Industrial Medicine. **29**: 602-608.

Hunter, S., Crome, P., Sim, J., Donaldson, C. and Pomeroy, V. M. (2006). Development of treatment schedules for research: a structured review to identify methodologies used and a worked example of 'mobilisation and tactile stimulation', for stroke patients. Physiotherapy. **92**: 195-207.

Hurn, J., Kneebone, I. and Cropley, M. (2006). Goal setting as an outcome measure: A systematic review. Clinical Rehabilitation. **20**(9): 756-772.

Hurvitz, E. A., Conti, G. E. and Brown, S. H. (2003). Changes in movement characteristics of the spastic upper extremity after botulinum toxin injection. Archives of Physical Medicine and Rehabilitation. **84**(3): 444-454.

INVOLVE (2009). Patient and public involvement in research and research ethics committee review. DoH. London, Department of Health.

Jackson, D., Horn, S., Kersten, P. and Turner-Stokes, L. (2006). Development of a pictorial scale of pain intensity for patients with communication impairments: initial validation in a general population Clinical Medicine. **6**: 580-585.

Jaeschke, R., Singer, J. and Guyatt, G. H. (1989). Measurement of health status. Ascertaining the minimal clinically important difference. Control Clinical Trials. **10**: 407-415.

Jahangir, A. W., Tan, H. J., Norlinah, M. I., Nafisah, W. Y., Ramesh, S., Hamidon, B. B. and Raymond, A. A. (2007). Intramuscular injection of botulinum toxin for the treatment of wrist and finger spasticity after stroke. Medical Journal of Malaysia. **62**(4): 319-322.

Jette, A. M. (2005). The post-stroke rehabilitation outcomes project. Archives of Physical Medicine and Rehabilitation. **86**(Supplement 2): S124-S125.

Jolliffe, I. T. (2002). Principal Component Analysis. New York, USA, Springer-Verlag.

Jones, D., Stephens, J., Rochester, L., Ashburn, A. and Stack, E. (2009). Service user and carer involvement in physiotherapy practice, education and research: getting involved for a change. NZ Journal of Physiotherapy. **37**(1): 29-35.

Jones, L. (1990). Jebson test of hand function (British Version). London, National Hospital for Neurology and Neurosurgery.

Jones, L., Lewis, Y., Harrison, J. and Wiles, C. M. (1996). The effectiveness of occupational therapy and physiotherapy in multiple sclerosis patients with ataxia of the upper limb and trunk. Clinical Rehabilitation. **10**(4): 277-282.

Joreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. Psychometrika. **36**: 109-133.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. Educational and Psychological Measurement. **20**: 141–151.

Kaji, R., Osako, Y., Suyama, K., Maeda, T., Uechi, Y. and Iwasaki, M. (2010). Botulinum toxin type A in post-stroke upper limb spasticity. Current Medical Research & Opinion. **8**: 1983-1992.

Kanovsky, P., Slawek, J., Denes, Z., Platz, T., Sassin, I., Comes, G. and Grafe, S. (2009). Efficacy and safety of botulinum neurotoxin NT 201 in post-stroke upper limb spasticity. Clinical Neuropharmacology. **5**: 259-265.

Katalinic, O. M., Harvey, L. A., Herbert, R. D., Moseley, A. M., Lannin, N. A. and Schurr, K. (2010). Stretch for the treatment and prevention of contractures. Cochrane Database of Systematic Reviews.(9): 1-29.

Keith, R. A., Granger, C. V., Hamilton, B. B. and Sherwin, F. S. (1987). The Functional Independence Measure; a new tool for rehabilitation. In Eisenberg, M. G. and Grzesiak, R. C. Advances in Clinical Rehabilitation. New York, Springer Publishing Company**:** 6-18.

Kiresuk, T. and Sherman, R. (1968). Goal attainment scaling: a general method of evaluating comprehensive mental health programmes. Community Mental Health Journal. **4**: 443-453.

Kiresuk, T., Smith, A. and Cardillo, J. (1994). Goal attainment scaling: application, theory and measurement. New York, Lawrence Erlbaum Associates.

Kitzinger, J. (1994). The methodology of focus groups: the importance of interaction between research participants. Sociology of Health and Illness. **16**: 103-121.

Kline, P. (1994). An easy guide to factor analysis. London, Routledge.

Kline, P. (2000). Rasch scaling and other scales. In. Handbook of Psychological Testing. London, Routledge**:** 70-96.

Kline, P. (2000b). The new psychometrics - Science, Psychology and Measurement. London, Routledge.

Kline, P. (2000c). Handbook of psychological testing. London and New York, Routledge; Taylor & Francis group.

Koh, C. L., Hsueh, I. P., Wang, W. C., Sheu, C. F., Yu, T. Y., Wang, B. S. and Hsieh, C. L. (2006). Validation of the action research arm test using item response theory in patients after stroke. Journal of Rehabilitation Medicine. **38**: 375-380.

Kong, K.-H., Neo, J.-J. and Chua, K. S. (2007). A randomised controlled study of botulinum toxin A in the treatment of hemiplegic shoulder pain associated with spasticity. Clinical Rehabilitation. **21**(1): 28-35.

Kopp, B., Kunkel, A., Flor, H., Platz, T., Rose, U., Mauritz, K., Gresser, K., McCulloch, K. L. and Taub, E. (1997). The arm motor ability test: Reliability, validity and sensitivity to change of an instrument for assessing disabilities in activities of daily living. Archives of Physical Medicine and Rehabilitation. **78**: 615-620.

Krantz, D. H., Luce, R. D., Suppes, P. and Tversky, A. (1971). Foundations in measurement. New York, Academic Press.

Kuder, G. F. and Richardson, M. W. (1937). The theory of estimation of test reliability Psychometrika. **2**: 151-160.

Lacey, G. and MacNamara, S. (2000). User involvement in the design and evaluation of a smart mobility aid. Journal of Rehabilitation Research & Development. **37**(6): 709-723.

Lagalla, G., Danni, M., Reiter, F., Ceravolo, M. G. and Provinciali, L. (2000). Post-stroke spasticity management with repeated botulinum toxin injections in the upper limb. American Journal of Physical Medicine & Rehabilitation. **79**(4): 377-384.

Lai, J., Francisco, G. and Willis, F. (2009). Dynamic splinting after treatment with botulinum toxin type-A: a randomized controlled pilot study. Advanced Therapy. **26**(2): 241-248.

Lance, J. W. (1980). Symposium synopsis. In Feldman, R. G., Young, R. R., Koella, W. P. and (editors). Spasticity: disordered motor control. Chicago:, Yearbook Medical Publishers**:** 465-494.

Landis, J. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics. **33**: 159-174.

Lannin, N., Cusick, A., McCluskey, A. and Herbert, R. D. (2007). Effects of splinting on wrist contracture after stroke; A randomised controlled trial. Stroke. **38**: 111-116.

Lannin, N. and Herbert, R. D. (2003). Is hand splinting effective for adults following stroke? A systematic review and methodological critique of published research. Clinical Rehabilitation. **17**: 807-816.

Latham, G. P. and Locke, E. A. (2007). New developments in and directions for goal-setting research. European Psychologist. **12**(4): 705-717.

Law, M. and Baum, C. (2001). Measurement in Occupational Therapy. In Law, M., Baum, C. and Dunn, W. Measuring Occupational Performance: Supporting Best Practice in Occupational Therapy, Thorofare, Slack Incorporated**:** 3-19.

Lawley, D. N. and Maxwell, A. E. (1963). Factor analysis as a statistical method. London, Butterworth.

Liang, M. H. (2000). Longitudinal construct validity: Establishment of clinical meaning in patient evaluation instruments. Medical Care. **38 (supplement II)**: S84-S90.

Linacre, J. M. (1994). Sample size and item calibration stability. Rasch Measurement Transactions. **7**: 328, At www.rasch.org/rmt/rmt74m.htm.

Lindberg, P., Schmitz, C., Forssberg, H., Engardt, M. and Borg, J. (2004). Effects of passive-active movement training on upper limb motor function and cortical activation in chronic patients with stroke: a pilot study. Journal of Rehabilitation Medicine. **36**(3): 117-123.

Loevinger, J. (1948). The techniques of homogeneous tests compared with some aspects of 'scale analysis' and factor analysis. Psychological Bulletin. **45**: 507-530.

Lomas, J., Pickard, L. and Mohide, A. (1987). Patient versus Clinician Item Generation for Quality-of-Life Measures: The Case of Language-Disabled Adults. Medical Care. **25**(8): 764-769.

Lord, F. M. (1952). A theory of test scores. New York, Psychometric Society.

Lord, F. M. (1974). Individualised Testing and Item Characteristic Curve Theory. Princeton, ETS.

Lord, F. M. and Novack, T. A. (1968). Statistical theories of mental test scores. Reading, Addison-Wesley.

Lorge, I. (1951). The fundamental nature of measurement. In Lindquist, F. Educational measurement. Washington DC, American Council of Education.

Luce, R. D. and Turkey, J. W. (1964). Simultaneous conjoint measurement: a new type of fundamental measurement. Journal of Mathematical Psychology. **1**: 1-27.

Lynch, H. T. and Lanspa, S. J. (2010). Colorectal cancer survival advantage in MUTYH-associated polyposis and Lynch syndrome families. Journal of the National Cancer Institute. **102**(22): 1687.

Malec, J. F. (1999). Goal attainment scaling in rehabilitation Neuropsychological Rehabilitation. **9**(3-4): 253-275.

Marco, E., Duarte, E., Vila, J., Tejero, M., Guillen, A. and Boza, R. (2007). Is botulinum toxin type A effective in the treatment of spastic shoulder pain in patients after stoke? A double-blind randomised clinical trial. Journal of Rehabilitation Medicine. **39**: 440-447.

Martin, M., Kosinski, M., Bjorner, J. B., Ware, J. E., MacLean, R. and Li, T. (2007). Item response theory methods can improve the measurement of physical function by combining the modified health assessment questionnaire and the SF-36 physical function scale. Quality of Life Research. **16**: 647-660.

Massof, R. W. (2005). Application of stochastic measurement models to visual function rating scale questionnaires. Ophthalmic Epidemiology. **12**: 103-124.

McCrory, P., Turner-Stokes, L., Baguley, I., De Graaff, S., Katrak, P., Sandanam, J., Davies, L., Munns, M. and Hughes, A. (2009). Botulinum toxin A for treatment of upper limb spasticity following stroke: A multi-centre randomized placebo-controlled study of the effects on quality of life and other person-centred outcomes. Journal of Rehabilitation Medicine. **41**: 536-544.

McDowell, I. and Newell, C. (1987). Measuring health - A guide to rating scales and questionnaires, Oxford University Press.

McHorney, C. A. and Tarlov, A. (1995). Individual-patient monitoring in clinical practice: Are available health status measures adequate? Quality of Life Research. **4**: 293-307.

McPherson, K., Berry, A. and Pentland, B. (1997). Relationships between cognitive impairments and functional performance after brain injury, as measured by the functional assessment measure (FIM + FAM). Neuropsychological Rehabilitation. **7**(3): 241-257.

Meadows, K. A., Twidale, F. and Rodgers, D. (1998). Action research - a model for introducing standardised health assessment in general practice: an exploratory study. Journal of Evaluation in Clinical Practice **4**(3): 225-229.

Medical Outcomes Trust (2002). Scientific Advisory Committee of the Medical Outcomes Trust; Assessing health status and quality-of-life instruments: attributes and review criteria. Quality of Life Research. **11**: 193-205.

MedStats.Org (2011). MedStats.Org/sens-spec.

Merriam-Webster Online Dictionary (2009). Merriam-Webster Online Dictionary., Available from: http://www.merriam-webster.com/dictionary (accessed 17th September 2009).

Messick, S. (1980). Test validity and the ethics of assessment. American Psychologist. **35**: 1012-1027.

Meythaler, J., Vogtle, L. and Brunner, R. (2009). A preliminary assessment of the benefits of the addition of botulinum toxin a to a conventional therapy program on the function of people with longstanding stroke. Archives of Physical Medicine & Rehabilitation. **90**: 1453-1461.

Michell, J. (1990). An introduction to the logic of psychological measurement. London, Lawrence Erlbaum associates.

Michell, J. (1999). Measurement in psychology: a critical history of a methodological concept. New York, Cambridge University Press.

Moe-Nilssen, R., Nordin, E. and Lundin-Olsson, L. (2008). Criteria for evaluation of measurement properties of clinical balance measures for use in fall prevention studies. Journal of Evaluation in Clinical Practice. **14**: 236-240.

Moher, D., Cook, D. J., Eastwood, S., Olkin, I., Rennie, D. and Stroup, D. F. (1999). Improving the quality of reports of meta-analysis of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. Lancet. **354**: 1896-1900.

Mokken, R. J. (1970). A theory and procedure of scale analysis. The Hague: Mouton.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P., Knol, D. L., M, B. L. and de Vet, H. C. W. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Quality of Life Research. **19**: 539-549.

Molenaar, I. W., Sijtsma, K. and Boer, P. (2000). Users Manual MSP5 for Windows; A program for Mokken scale analysis for polytomous items, iec ProGAMMA, Groningen, Netherlands.

Moseley, A. M., Hassett, L. M., Leung, J., Clare, J. S., Herbert, R. D. and Harvey, L. A. (2008). Serial casting versus positioning for the treatment of elbow contractures in adults with traumatic brain injury: a randomized controlled trial. Clinical Rehabilitation. **22**: 406-417.

Nagel, E. (1931). "Measurement", Erkenntnis, Volume 2, Number 1, Springer, the Netherlands.

Nakayama, H., Jorgensen, H. S., Raaschou, H. O. and Olsen, T. S. (1994). Recovery of upper limb extremity function in stroke patients: the Copenhagen stroke study. Archives of Physical Medicine and Rehabilitation. **75**(4): 394-398.

Nevo, B. (1985). Face validity revisited. Journal of Educational Measurement. **22**: 287-293.

Nijsten, T., Unaeze, J. and Stern, R. S. (2006). Refinement and reduction of the Impact of Psoriasis Questionnaire: Classical Test Theory vs. Rasch analysis. Epidemiology and Health Services Research. **154**: 692-700.

Norman, G. and Streiner, D. (2000). Biostatistics: The bare essentials (2nd Edition). Toronto, B C Decker.

Norman, G. R., Sloan, J. A. and Wyrwich, K. W. (2003). Interpretation of changes on health related quality of life. The remarkable universality of half a standard deviation. Medical Care. **41**: 582-592.

Nunnally, J. (1970). Introduction to psychological measurement. New York, USA, McGraw-Hill.

Nunnally, J. (1978). Psychometric Theory. New York, McGraw-Hill.

Nunnally, J. and Bernstein, I. H. (1994). Psychometric Theory. New York, USA, McGraw-Hill.

O'Rourke, N. and Tuokko, H. A. (2003). Psychometric properties of an abridged version of the Zarit Burden Interview within a representative Canadian caregiver sample. The Gerontologist. **43**: 121-127.

Osborne, J. W. and Costello, A. B. (2004). Sample size and subject to item ratio in principal components analysis. Practical Assessment, Research & Evaluation. **9**(11): Retrieved January 4, 2011, http://PAREonline.net/getvn.asp?v=9&n=11.

Page, S. and Levine, P. (2003). Forced use after TBI: promoting plasticity and function through practice. Brain Injury. **17**(8): 675-684.

Page, S. J., Elovic, E., Levine, P. and Sisto, S. A. (2003). Modified constraint-induced therapy and botulinum toxin A: a promising combination. American Journal of Physical Medicine & Rehabilitation. **82**(1): 76-80.

Pallant, J. F. and Tennant, A. (2006). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). British Journal of Clinical Psychology. **76**(4): 781-801.

Pandyan, A. D., Gregoric, M., Barnes, M. P., Wood, D., van Wijck, F., Burridge, J., Hermens, H. and Johnson, G. R. (2005). Spasticity: clinical perceptions,

neurological realities and meaningful measurement. Disability and Rehabilitation. **27**: 2-6.

Parry, R. H., Lincoln, N. B. and Vass, C. D. (1999). Effect of severity of arm impairment on response to additional physiotherapy early after stroke. Clinical Rehabilitation. **13**(3): 187-198.

Pedhazur, E. J. (1997). Multiple regression in behavioural research: Explanation and Prediction. Fort Worth, USA, Harcourt Brace College Publishers.

Penta, M., Tesio, L., Arnould, C., Zancan, A. and Thonnard, J.-L. (2001). The ABILHAND Questionnaire as a Measure of Manual Ability in Chronic Stroke Patients - Rasch-Based Validation and Relationship to Upper Limb Impairment. Stroke. **32**: 1627-1634.

Penta, M., Thonnard, J.-L. and Tesio, L. (1998). ABILHAND: A Rasch-Built Measure of Manual Ability. Archives of Physical Medicine and Rehabilitation. **79**: 1038-1042.

Perline, R., Wright, B. D. and Wainer, H. (1979). The Rasch model as additive conjoint measurement. Applied Psychological Measurement. **3**(2): 237-255.

Playford, D. (2008). Outcome measurement in neurological disease. Current Opinion in Neurology. **21**: 649-653.

Playford, D., Siegert, R., Levack, W. and Freeman, J. A. (2009). Areas of consensus in rehabilitation: a conference report. Clinical Rehabilitation. **23**: 334-344.

Poissanta, L. and Mayob, N. E. (2004). The use of the International Classification of Functioning, Disability and Health (ICF) to create a clinical problem list in an Electronic Health Record (EHR). MEDINFO. 1813.

Pollard, B., Dixon, D., Dieppe, P. and Johnston, M. (2009). Measuring the ICF components of impairment, activity limitation and participation restriction: an item analysis using classical test theory and item response theory. Health and Quality of Life Outcomes. **7**(41): 1-20.

Pope, P. M. (2002). Postural management and special seating. In Edwards, S. Neurological Physiotherapy - a problem solving approach. . London, Churchill Livingstone.

Popovic, M. B., Popovic, D. B., Sinkjaer, T., Stefanovic, A. and Schwirtlich, L. (2003). Clinical evaluation of functional electrical therapy in acute hemiplegic subjects. Journal of Rehabilitation Research and Development. **40**(5): 443-453.

Powell, C. (2003). The Delphi technique: Myths and realities. Journal of Advanced Nursing. **41**: 376-382.

Raine, S. (2006). Defining the Bobath concept using the Delphi technique. Physiotherapy Research International. **11**: 4-11.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Chicago: MESA.

Raykov, T. and Marcoulides, G. A. (2011). Introduction to Psychometric Theory, Routledge Taylor & Francis Group, London.

Reed, J. and Roskell-Payton, V. (1997). Focus groups: issues of analysis and interpretation. Journal of Advanced Nursing. **26**: 765-771.

Reeve, B. B. (2006). An Introduction to Modern Measurement Theory, U.S. National Institutes of Health - Division of Cancer Control and Population Sciences. www.cancer.gov.

Reid, N. (1988). The Delphi Technique: its contribution to the evaluation of professional practice. In Ellis, R. Professional competence and quality assurance in the caring professions. London, Chapman Hall.

Richardson, D., Edwards, S., Sheean, G. L., Greenwood, R. J. and Thompson, A. J. (1997). The effect of botulinum toxin on hand function after incomplete spinal cord injury at the level of C5/6: a case report. Clinical Rehabilitation. **11**(4): 288-292.

Richardson, D., Greenwood, R., Sheean, G., Thompson, A. and Edwards, S. (2000). Treatment of focal spasticity with botulinum toxin: effect on the 'positive support reaction'... including commentary by Burridge J. Physiotherapy Research International. **5**(1): 62-72.

Ring, H. and Rosenthal, N. (2005). Distributed form of constraint-induced movement therapy improves functional outcome and quality of life after stroke. Archives of Physical Medicine and Rehabilitation. **86**(2): 204-9.

Rockwood, K., Joyce, B. and Stolee, P. (1997). Use of goal attainment scaling in measuring clinically important change in cognitive rehabilitation patients. Journal of Clinical Epidemiology. **50**(5): 581-588.

Rodgers, H. (2008). What is the clinical effect and cost effectiveness of treating upper limb spasticity due to stroke with botulinum toxin? London, UK, National Research Register (Health Technologies Assessment).

Roorda, L. D., Roebroeck, M. E., Lankhorst, G. J., Van Tilburg, T. and Bouter, L. M. (1996). Measuring functional limitations in rising and sitting down: development of a questionnaire Archives of Physical Medicine and Rehabilitation. **77**: 663-669.

Roskam, E. E. (1985). Current issues in item-response theory: beyond psychometrics In Roskam, E. E. Measurement and personality assessment. Amsterdam, Elsevier.

Rowland, T. J. and Gustafsson, L. (2008). Assessment of upper limb ability following stroke: a review. British Journal of Occupational Therapy. **71**(10): 427-437.

Royal College of Physicians (2002). Guidelines for the use of botulinum toxin (BTX) in the management of spasticity in adults. London, UK, Royal College of Physicians, Clinical Effectiveness and Evaluation Unit.

Royal College of Physicians, British Society of Rehabilitation Medicine, Chartered Society of Physiotherapy and Association of Chartered Physiotherapist Interested in Neurology (2009). Spasticity in adults: management using botulinum toxin - National Guidelines. London, Royal College of Physicians, Clinical Effectiveness and Evaluation Unit.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph. **17**.

Scarpelli, A. C., Paiva, S. M., Pordeus, I. A., Ramos-Jorge, M. L., Varni, J. W. and Allison, P. J. (2008). Measurement properties of the Brazilian version of the Paediatric Quality of Life Inventory (PedsQL) cancer module scale. Health and Quality of Life Outcomes. **6**(7): 1-11.

Schmitt, N. (1996). Uses and abuses of Coefficient Alpha. Psychological Assessment. **8**(4): 350-353.

Schwartz, C. E. and Sprangers, M. A. G. (1999). Methodological approaches for assessing response shift in longitudinal health related quality of life research. Social Science and Medicine. **48**: 1531-1548.

Shaw, L., Rodgers, H., Price, C., Van Wijck, F., Shackley, P., Steen, N., Barnes, M., Ford, G., Graham, L. and BoTULS_investigators (2010). BoTULS: a multicentre randomised-controlled trial to evaluate the clinical effectiveness and cost-effectiveness of treating upper limb spasticity due to stroke with botulinum toxin A. Health Technologies Assessment. **14**(26): 1-113, iii-iv.

Sheean, G., Lannin, N. A., Turner-Stokes, L., Rawicki, B. and Snow, B. J. (2010). Botulinum toxin assessment, intervention and after-care for upper limb

hypertonicity in adults: international consensus statement. European Journal of Neurology. **17**(Supplement 2): 74-93.

Sheean, G. L. (2001). Botulinum treatment of spasticity: Why is it difficult to show a functional benefit? Trauma and Rehabilitation. 771-776.

Siegert, R., Jackson, D., Tennant, A. and Turner-Stokes, L. (2010). Factor analysis and Rasch analysis of the Zarit Burden interview for acquired brain injury carer research. Journal of Rehabilitation Medicine. **42**: 302-309.

Sim, J. and Snell, J. (1996). Focus groups in physiotherapy: Evaluation and research. Physiotherapy. **82**(3): 189-198.

Simpson, D., Gracies, J., Yablon, S., Barbano, R. and Brashear, A. (2009). Botulinum neurotoxin versus tizanidine in upper limb spasticity: a placebo-controlled study. Journal of Neurology, Neurosurgery & Psychiatry. **80**: 380-385.

Simpson, D. M., Alexander, D. N., O. Brien, C. F., Tagliati, M., Aswad, A. S., Leon, J. M., Gibson, J., Mordaunt, J. M. and Monaghan, E. P. (1996). Botulinum toxin type A in the treatment of upper extremity spasticity: a randomised, double-blind, placebo-controlled trial. Neurology. **46**(5): 1306-1310.

Slade, M. (2002). The use of patient-level outcomes to inform treatment. Epidemiologiae Psichiatria Sociale. **11**(1): 20-27.

Slade, M. (2002b). Routine outcome assessment in mental health services. Psychological Medicine. **32**: 1339-1343.

Slade, M., Thornicroft, G. and Glover, G. (1999). The feasibility of routine outcome measures in mental health. Social Psychiatry and Psychiatric Epidemiology. **34**: 243-249.

Smith, A. B., Wright, P., Selby, P. J. and Velikova, G. (2007). A Rasch and Factor Analysis of the Functional Assessment of Cancer Therapy-General (FACT-G). Health and Quality of Life Outcomes. **5**(19).

Smith, J. A., Flowers, P. and Osborn, M. (1997). Interpretative phenomenological analysis and the psychology of health and illness. In Yardley, L. Material Discourses of Health and Illness. London, Routledge**:** 68-91.

Smith, S. J., Ellis, E., White, S. and Moore, A. P. (2000). A double-blind placebo-controlled study of botulinum toxin in upper limb spasticity after stroke or head injury. Clinical Rehabilitation. **14**(1): 5-13.

Sokal, M. (1971). The unpublished autobiography of James McKeen Cattell. American Psychologist. **26**(7): 626-635.

Sommerfeld, D. K., Elsy, U. B., Svensson, A. K., Holmqvist, L. W. and von Arbin, M. H. (2004). Spasticity after stroke Its occurrence and association with motor impairments and activity limitations. Stroke. **35**: 134-140.

Spearman, C. (1904). General intelligence, objectively determined and measured. American Journal of Psychology. **15**: 201-93.

SPSS (2000). Statistical package for the social sciences. Chicago, SPSS.

Stata (2001). Stata statistical software. College Station, Stata Corporation.

Stevens, S. S. (1946). On the theory of scales of measurement. Science. **103**: 677-680.

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In Stevens, S. S. Handbook of experimental psychology. New York, Wiley**:** 1-49.

Stevens, S. S. (1959). Measurement, psychophysics and utility. In Churchman, C. W. and Ratoosh, P. Measurement: definition and theories. New York, Wiley**:** 18-63.

Stevenson, V. L. and Jarrett, L. (2006). Spasticity management - A practical multidisciplinary guide. London, Informa healthcare.

Stolee, P., Rockwood, K., Fox, R. A. and Streiner, D. L. (1992). The use of goal attainment scaling in a geriatric care setting. Journal of the American Geriatrics Society. **40**(6): 574-578.

Stolee, P., Zaza, C., Pedlar, A. and Myers, A. M. (1999). Clinical experience with Goal Attainment Scaling in geriatric care. Journal of Aging & Health. **11**(1): 96-124.

Stratford, P. (1989). Reliability: consistency or differentiating among subjects? Physical Therapy. **69**: 299-300.

Strauss, J. H. and Ziegler, L. H. (1975). The Delphi technique and its uses in social science research. Journal of Creative Behaviour. **9**(4): 253-259.

Streiner, D. (2003a). Starting at the beginning: An introduction to coefficient Alpha and internal consistency. Journal of Personality Assessment. **80**(1): 99-103.

Streiner, D. and Norman, G. (2003). Health measurement scales; a practical guide to their development and use, Oxford University Press.

Streiner, D. L. (2003b). Clinimetrics vs. psychometrics: an unnecessary distinction.[comment]. Journal of Clinical Epidemiology. **56**(12): 1142-1145.

Stucki, G., Cieza, A. and Melvin, J. (2007). The international classification of functioning, disability and health: a unifying model for the conceptual description of the rehabilitation strategy. Journal of Rehabilitation Medicine (39): 279-285.

Sun, S., Hsu, C., Sun, H., Hwang, C., Yang, C. and Wang, J. (2010). Combined botulinum toxin type A with modified constraint-induced movement therapy for chronic stroke patients with upper extremity spasticity: a randomized controlled study. Neurorehabilitation & Neural Repair. **24**(1): 34-41.

Tabachnick, B. G. and Fidell, L. S. (2001). Using multivariate statistics. New York, Harper Collins.

Takahashi, T. (2004). ICF illustrated library International University of Health and Welfare and TAI Human Research. Inc. (online) Available from: http://www.icfillustration.com/top_e.html (accessed 3rd March 2009).

Tamber, A.-L., Wilhelmsen, K. T. and Strand, L. I. (2009). measurement properties of the Dizziness handicap Inventory by cross-sectional and longitudinal designs. Health and Quality of Life Outcomes. **7**(101): 1-16.

Tansella, M. and Thornicroft, G. (2001). Mental health outcome measures, Springer-Verlag, Berlin.

Taub, E., Miller, N. E., Novack, T. A., Cook, E. W., Fleming, W. C., Nepomuceno, C. S., Connell, J. S. and Crago, J. E. (1993). Technique to improve chronic motor deficit after stroke. Archives of Physical Medicine and Rehabilitation. **74**: 347-354.

Teasdale, G. and Jennett, B. (1974). Assessment of coma and impaired consciousness. A practical scale. Lancet **2**(7872): 81–4.

Tennant, A. (2007). Goal attainment scaling: Current methodological challenges. Disability and Rehabilitation. **29**: 1583-1588.

Tennant, A. and Conaghan, P. G. (2007b). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Care and Research. **57**(8): 1358-1362.

Tennant, A. and Young, C. (1997). Coma to community: continuity in measurement... First International Outcome Measurement Conference, co-sponsored by Rehabilitation Foundation, Inc., and the MESA Psychometric Laboratory at the University of Chicago. Physical Medicine and Rehabilitation State of the Art Reviews. **11**(2): 375-384.

Teo, Y. Y. and Chong, F. F. (2006). On the usage of principle components analysis and multiple testing. American Journal of Respiratory and Critical Care Medicine. **173**: 574.

Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J. H., Bouter, L. M. and de Vet, H. C. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. Journal of Clinical Epidemiology. **60**(1): 34-42.

Terwee, C. B., Mokkink, L. B., Steultjens, M. P. and Dekker, J. (2006). Performance-based methods for measuring the physical function of patients with osteoarthritis of the hip or knee: a systematic review of measurement properties. Rheumatology. **45**(7): 890-902.

Thissen, D., Reeve, B. B., Bjorner, J. B. and Chang, C.-H. (2007). Methodological issues for building item banks and computerized adaptive scales. Quality of Life Research. **16**: 109-119.

Thissen, D. and Steinberg, L. (1984). A response model for multiple choice items. Psychometrika. **51**: 567-577.

Thompson, A., Jarrett, L., Lockley, L., Marsden, J. and Stevenson, V. L. (2005). Clinical management of spasticity. Journal of Neurology, Neurosurgery and Psychiatry. **76**: 459-463.

Thomson, W. (1891). Popular lectures and addresses. London, Macmillan.

Thurstone, L. L. (1927). A law of comparative judgement. Psychological Review. **34**: 278-286.

Thurstone, L. L. (1947). Multiple factor analysis Chicago, University of Chicago Press.

Turner-Stokes, L. (2009a). Upper limb intervention for spasticity (ULIS) - Botulinum toxin-A cohort survey. London, King's College London**:** 1-34.

Turner-Stokes, L. (2009b). Goal attainment scaling (GAS) a practical guide. Clinical Rehabilitation. **23**(4): 362-370.

Turner-Stokes, L., Baguley, I., De Graaff, S., Katrak, P., Davies, L., McCrory, P. and Hughes, A. (2010). Goal attainment scaling in the evaluation of treatment of upper limb spasticity with botulinum toxin: A secondary analysis from a double-blind placebo-controlled randomised clinical trial. Journal of Rehabilitation Medicine. **42**: 81-89.

Turner-Stokes, L. and Jackson, D. (2006). Assessment of shoulder pain in hemiplegia: sensitivity of the ShoulderQ. Disability and Rehabilitation. **28**(6): 389-395.

Turner-Stokes, L., Nyein, K., Turner-Stokes, T. and Gatehouse, C. (1999). The UK FIM+FAM: development and evaluation. Functional Assessment Measure. Clinical Rehabilitation. **13**(4): 277-287.

Tversky, A. and Gati, I. (1982). Similarity, separability and the triangle inequality. Psychological Review. **89**: 123-154.

Tyson, S. F. and Kent, R. M. (2010). Orthotic devices after stroke and other nonprogressive brain lesions. Stroke, American Heart Association Inc. **41:** 00-00.

Uswatte, G. and Taub, E. (2005). Implications of the learned non-use formulation for measuring rehabilitation outcomes: lessons from constraint-induced movement therapy. Rehabilitation Psychology. **50**: 34-42.

Uswatte, G., Taub, E., Morris, D., Light, K. and Thompson, P. (2006). The Motor Activity Log-28 Assessing daily use of the hemiparetic arm after stroke. Neurology. **67**: 1189-1194.

Uswatte, G., Taub, E., Morris, D., Vignolo, M. and McCulloch, K. L. (2005). Reliability and validity of the upper-extremity Motor Activity Log-14 for measuring real world arm use. Stroke. **36**: 2493-2496.

van Abswoude, A. A. H., Vermunt, J. K., Hemker, B. T. and van der Ark, L. A. (2004). Mokken Scale Analysis Using Hierarchical Clustering Procedures. Applied Psychological Measurement. **28**(5): 332-354.

van de Ven-Stevens, L. A., Munneke, M., Terwee, C. B., Spauwen, P. H. and van der Linde, H. (2009). Clinimetric properties of instruments to assess activities in patients with hand injury: a systematic review of the literature. Archives of Physical Medicine & Rehabilitation. **90**(1): 151-69.

Van de Winckel, A., Feys, H., van der Knaap, S., Messerli, R., Baronti, F., Lehmann, R., Van Hemelrijk, B., Pante, F., Perfetti, C. and De Weerdt, W. (2006). Can quality of movement be measured? Rasch analysis and inter-rater reliability of the motor evaluation scale for upper extremity in stroke patients (MESUPES). Clinical Rehabilitation. **20**: 871-884.

van der Lee, J. H., Beckerman, H., Knol, D. L., de Vet, H. C. V. and Bouter, L. M. (2004). Clinimetric properties of the Motor Activity Log for the assessment of arm use in hemiparetic patients. Stroke. **35**: 1-5.

Van der Lee, J. H., Roorda, L. D., Beckerman, H., Lankhorst, G. J. and Bouter, L. M. (2002). Improving the action research arm test: a unidimensional hierarchical scale. Clinical Rehabilitation. **16**: 646-653.

van der Putten, A., Vlaskamp, C., Reynders, K. and Nakken, H. (2005). Movement skill assessment in children with profound multiple disabilities: a psychometric

analysis of the top down motor milestone test. Clinical Rehabilitation. **19**: 635-643.

van Kuijk, A. A., Geurts, A. C., Bevaart, B. J. and van Limbeek, J. (2002). Treatment of upper extremity spasticity in stroke patients by focal neuronal or neuromuscular blockade: a systematic review of the literature. Journal of Rehabilitation Medicine. **34**(2): 51-61.

van Schuur, W. H. (2003). Mokken scale analysis: Between the Guttman scale and parametric item response theory. Political Analysis. **11**: 139-163.

Vattanaslip, W., Ada, L. and Crosbie, J. (2000). Contribution of thixotropy, spasticity and contracture to ankle stiffness after stroke. . Journal of Neurology Neurosurgery and Psychiatry. **69**: 34-39.

Wade, D. (1992a). Evaluating outcome in stroke rehabilitation. Scandinavian Journal of Rehabilitation Medicine. **Supplement 26**: 97-104.

Wade, D. (2009). Goal setting in rehabilitation: an overview of what, why and how. Clinical Rehabilitation. **23**: 291-295.

Wade, D. T. (1992b). Measurement in neurological rehabilitation. Oxford, Oxford University Press.

Wade, D. T. and Collin, C. (1988). The Barthel ADL index: a standard measure of physical disability? International Disability Studies. **10**: 64-67.

Wade, D. T., Langton-Hewer, R., Wood, V., Skilbeck, C. E. and Ismail, I. M. (1983). The hemiplegic arm after stroke: measurement and recovery. Journal of Neurology, Neurosurgery and Psychiatry. **46**: 521-524.

Watkins, C. L., Leathley, M. J., Gregson, J. M., Moore, A. P., Smith, T. L. and Sharma, A. K. (2002). Prevalence of spasticity post stroke. Clinical Rehabilitation. **16**(5): 515-522.

WHO (1980). The International Classification of Impairments, Disabilities and Handicaps (ICIDH) - a manual of classification relating to the consequences of disease. Geneva, World Health Organisation.

WHO (2002). International Classification of Functioning, Disability and Health. Geneva, World Health Organisation.

Wiener-Ehrlich, W. R. (1978). Dimensional and metric structures in multidimensional stimuli. Perception and Psychophysics. **24**: 399-414.

Wiley, D. E., Schmidt, W. H. and Bramble, W. J. (1973). Studies of a class of covariance structure models. Journal of the American Statistical Society. **86**: 317-321.

Wismeijer, A. A. J., Sijtsma, K., van Assen, A. L. M. and Vingerhoets, J. J. M. (2008). A comparative study of the dimensionality of the self-concealment scale using principal component analysis and Mokken scale analysis. Journal of Personality Assessment. **90**(4): 323-334.

Wissel, J., Ward, A. B., Erztgaard, P., Bensmail, D., Heckt, M. J., Lejeune, T. M. and Schnider, P. (2009). European consensus table on the use of botulinum toxin type a in adult spasticity. Journal of Rehabilitation Medicine. **41**: 13-25.

Wolf, S. L., Catlin, P. A., Ellis, M., Archer, A. L., Morgan, B. and Piacentino, A. (2001). Assessing Wolf motor function test as outcome measure for research in patients after stroke. Stroke. **32**: 1635-1639.

Wright, B. D. and Tennant, A. (1996). Sample size again. Rasch Measurement Transactions. **9**: 468.

Wright, J. G. and Feinstein, A. R. (1992). A comparative contrast of clinimetric and psychometric methods for constructing indexes and rating scales. Journal of Clinical Epidemiology. **45**(11): 1201-1218.

Yablon, S. A., Agana, B. T., Ivanhoe, C. B. and Boake, C. (1996). Botulinum toxin in severe upper extremity spasticity among patients with traumatic brain injury: An open-labelled trial. Neurology. **47**(4): 939-944.

Yelnik, A. P. (2004). Pharmacology and upper limb post-stroke spasticity: A Review. [French]. Annales de Readaptation et de Medecine Physique. **47**(8): 575-589.

Yelnik, A. P., Colle, F. M., Bonan, I. V. and Vicaut, E. (2007). Treatment of shoulder pain in spastic hemiplegia by reducing spasticity of the subscapular muscle: a randomised, double-blind, placebo-controlled study of botulinum toxin A. Journal of Neurology, Neurosurgery and Psychiatry. **78**(8): 845-848.

Zaza, C., Stolee, P. and Prkachin, K. (1999). The application of goal attainment scaling in chronic pain settings. Journal of Pain and Symptom Management. **17**(1): 55-64.

Zwinkels, A., Geusgens, C., van de Sande, P. and van Heugten, C. (2004). Assessment of apraxia: inter-rater reliability of a new apraxia test, association between apraxia and other cognitive deficits and prevalence of apraxia in a rehabilitation setting. Clinical Rehabilitation. **18**: 819-827.